Development of a metadatabase to support a multi-organisational, multi-disciplinary river ecosystem research and management initiative – Experiences from the Kruger National Park Rivers Research Programme

HC Biggs[1], Y Coetzee[2]* and MC Dent[3]
[1] P/Bag X402, Skukuza, Kruger National Park, 1350, Mpumalanga, South Africa
[2] c/o EM Coetzee. P/Bag X402, Skukuza, Kruger National Park, 1350, Mpumalanga, South Africa
[3] Computing Centre for Water Research, University of Natal, P/Bag X01, Scottsville 3209, Pietermaritzburg, South Africa

Abstract

This paper describes the products and processes developed by the Information Systems Development and Management Subprogramme (ISDM) of the Kruger National Park Rivers Research Programme (KNPRRP) Phase II. The goals of the KNPRRP are to:

- inform researchers, system managers and stakeholders about the water quality and quantity requirements to sustain the natural environment of rivers which flow through the Kruger National Park; and
- develop, test and refine methods for predicting the response of the natural environments of rivers flowing through the Kruger National Park and in Southern Africa to changing water quality and patterns of supply.

A primary task in a programme of this nature is to gather and organise the metadata of interest. In the case of the multi-organisational, multi-disciplinary KNPRRP this was a task so large that it required the devotion of a separate subprogramme within the programme structure. The purpose of the ISDM subprogramme was to provide an information management system which enables the efficient capture, storage, retrieval and dissemination of the information to serve the metadata needs of researchers, decision makers and stakeholders.

The choice of personalities and processes employed to gather the information was central to the success of the project. In addition to providing a "hard facts" description of the methodology and achievements of ISDM, this paper explores some of the philosophical, psychological and business issues embodied in the choice of personalities and processes required for such a task.

The metadata catalogue is available through the world-wide web via the KNPRRP home page (http://www.ccwr.ac.za/knprrp). The information is presented in four broad categories, namely:

- documents (both conventional publications and "grey" literature)
- digital data (including GIS and time-dependent climatic data)
- simulation models
- collections (including museum collections, videos, and photographs).

Within the KNPRRP the ISDM both served and derived inspiration and information from the KNPRRP's other three subprogrammes namely, the Decision Support System Development and Management Subprogramme (DSSDM), the Research Development and Management Subprogramme (RDM) and the Training and Information Technology Transfer Subprogramme (TITT).

Introduction

The following abbreviations are used regularly in this paper:

- Computing Centre for Water Research (CCWR)
- Decision Support System Development and Management Subprogramme (DSSDM)
- Department of Water Affairs and Forestry (DWAF)
- Geographical Information Systems (GIS)
- Information Systems Development and Management Subprogramme (ISDM)
- Kruger National Park Rivers Research Programme (KNPRRP)
- Research Development and Management Subprogramme (RDM)
- South African National Parks (SANP)
- Training and Information Technology Transfer Subprogramme (TITT)
- Water Research Commission (WRC)

* To whom all correspondence should be addressed.
☎(013) 735-5039; fax (013) 735-5039; e-mail: yvedan02@mweb.co.za

The KNPRRP is a multi-disciplinary, multi-organisational research programme which has the following goals:

- To inform researchers, system managers and stakeholders about the water quality and quantity requirements to sustain the natural environment of rivers which flow through the Kruger National Park.
- To develop, test and refine methods for predicting the response of the natural environments of rivers flowing through the Kruger National Park and in Southern Africa to changing water quality and patterns of supply.

A primary task in a programme of this nature is to gather and organise the metadata of interest. Metadata is a term used to describe data which describe datasets. In the case of the multi-organisational, multi-disciplinary KNPRRP this was a task so large that it required the devotion of a separate subprogramme within the programme structure. The purpose of ISDM was to provide an information management system which enables the efficient capture, storage, retrieval and dissemination of the information to

serve the metadata needs of researchers, decision-makers and stakeholders.

It is in the nature of such programmes that the researchers need access to a wide range of data and information. Such data and information are collected and stored by many different individuals at a number of institutions countrywide, which makes it difficult for individual researchers and managers to know about it. Accessing that information raises further difficulties. The KNPRRP considered it essential to know about and have access to these data and information and therefore committed one of its four subprogrammes to seek out and catalogue metadata pertaining to it.

The tasks of ISDM included collation, determining criteria, reviewing existing data management systems, developing guidelines, focusing projects on data requirements, developing intellectual property guidelines and identifying strategic gaps. An extensive catalogue was developed. It contained:

- documents published (both conventional and "grey" literature)
- digital data (including GIS and time-series climatic and hydrological data)
- simulation models
- collections (including museum collections, videos, and photographs).

As an adjunct, selected documents were abstracted, and an extensive series of fixed-point river photographs digitised.

Business issues

The management of research is becoming increasingly business-like in its approach, in response to the difficult economic climate. It is therefore considered appropriate to share some of the approaches to the research business issues which were encountered.

Personnel selection

The senior author was selected by the KNPRRP's Senior Management at least partly because of his knowledge of a wide spectrum of rivers projects, gained during the course of his work as a datametric support professional in the Scientific Services Division of the SANP in Skukuza. The unit of which he was at the time the manager, was already acting as a de-facto "information-sharing node" in the region.

The information collators were selected according to their knowledge of river initiatives, their scientific background and their potential to develop interviewing skills. One collator was based in the actual study area (where a large percentage of datasets could be traced) and the other one in Gauteng (where the large government departments holding national datasets, and where the closest universities involved in the KNPRRP, were based).

Approaching data custodians

All individuals who were considered as potential custodians of datasets were sent a formal letter requesting their participation in the KNPRRP, in which the latter was explained briefly to them and in which they were formally requested to participate. The actual collection of metadata was carried out through a mechanism of personal interview by the carefully selected collators.

Features of the data catalogue

A broad definition of data was taken, including documents (published and grey literature e.g. selected unpublished reports), digital datasets including geographical information systems, GIS layers, hard-copy maps (especially where these were not yet digitised), models and selected regularly-used model outputs, procedures and protocols, collections (such as museum collections) and videos.

Level of detail

The metadata entries were made as detailed as possible. The principle was to provide users with as much information as might be useful, without replicating the actual data. Data fields therefore also included unusual descriptors such as funding agencies and methods of access. A typical entry for a digital dataset is shown in Fig. 1.

Data classification system

It was decided to use the same classification system as the RDM subprogramme had employed for their research project classification scheme. This system was also considered most useful for stakeholders. It was used as the basic outline for the hard-copy and digital document versions of the metadata catalogue. The latter also included cross-referencing and a keyword scheme.

Relational database

A relational database version, with a versatile custom-built search engine, was also provided, and all products also placed on a world-wide website (http:/www.ccwr.ac.za/knprrp) at the CCWR in Pietermaritzburg, South Africa.

Updating the data catalogue

Ongoing additions to the catalogue were supported by automation procedures which could at any time produce an up-to-date desktop-published correctly-indexed version of the hard-copy or equivalent digital document.

Geographic extent of the catalogue information

Catalogue entries were restricted to those relevant to the aims of the programme in the catchments of the specified rivers. These extended downstream into Mozambique where possible. The upper part of the Olifants River, west of the escarpment, was excluded from the Phase II catalogue for reasons pertaining to the collection logistics. It was planned to incorporate these in a follow-up project. Entries of a very general nature (for example, a protocol used nationally and thus also in the KNPRRP) were included if considered locally important.

Metadata products

The metadata covered a wide range of topics as listed below.
- Biota including riparian vegetation, invertebrates, fish and other
- Hydrology and catchment studies
- Geomorphology and hydraulics
- GIS
- Water quality

```
┌─────────────────────────────────────────────────────────────────────────────┐
│                                                                               │
│  DIGITAL                    BIORIP-DIG-005                                     │
│                                                                               │
│               The riparian vegetation of the Sabie River, KNP.                │
│                                                                               │
│  Custodian: Bredenkamp G J                                                    │
│                                                                               │
│  Objective: Vegetation survey; floristic spp. data                            │
│                                                                               │
│  Period: December 1991 - February 1992                                        │
│                                                                               │
│  Keywords: riparian vegetation; Sabie River                                   │
│                                                                               │
│  Format: University of Pretoria mainframe (same format as used by the         │
│  Agricultural Research Council)                                               │
│                                                                               │
│  Fields: plot no; plant spp.; estimated % cover (approximately 100 spp.       │
│  per plot)                                                                     │
│                                                                               │
│  Missing values: N/A                                                          │
│                                                                               │
│  Accuracy: Approximately 100 sample plots                                     │
│                                                                               │
│  Usages: To calculate an index of vegetation biodiversity; to propose as      │
│  management zones; can provide the basis for site selection for further       │
│  research                                                                     │
│                                                                               │
│  Limitations: Single once-off survey done in summer                           │
│                                                                               │
│  Interchange: ASCII/generated as Braun-Blanquet plots                         │
│                                                                               │
│  Conjunctive DB: Document; original 1 : 10 000 or 1 : 8 000 maps showing      │
│  the sampling site positions at Skukuza; aerial photographs at Skukuza        │
│                                                                               │
│  Collection form: Follow the methodology implemented in this study for        │
│  further qualitative surveys                                                  │
│                                                                               │
│  Xreference: BIORIP-DIG-002;BIORIP-DIG-006;BIORIP-DOC-013;INF-DOC-002;        │
│  Gis-GIS-003;Gis-GIS-002;BIORIP-DIG-008;BIORIP-COL-002                        │
│                                                                               │
│  Funder: National Research Foundation (NRF)                                   │
│                                                                               │
│  Access: Author, with permission of NRF                                       │
│                                                                               │
│  O'Keeffe Classification No.  BIORIP-DIG-005                                   │
│                                                                               │
└─────────────────────────────────────────────────────────────────────────────┘
```

Figure 1
A typical entry (for example a digital record) is shown above

- Instream flow
- Conservation status and importance
- Legal/decision support systems/social issues.

Entries under each of these headings were then subgrouped into four major categories for storage, namely published documents, digital data, simulation models and collections. Currently a few hard copies of the catalogue are provided to key agencies and libraries, but general users are provided with digital documents only – if they wish to print these they can do so in their own organisation. The underlying data files are freely available on request.

The digital database version is no longer routinely supplied because it is felt that the search engines now available on the worldwide web enable easy searching of this database at site (http://www.ccwr.ac.za/knprrp). The source data on this site are regularly updated from the master copy of the database. The metadata page on the site was being visited about 30 times per month in 1998.

Documents (both conventional publications and "grey" literature)

Inclusion of grey literature expanded the usefulness of the catalogue, as unpublished reports from different agencies can be very difficult for users to trace. Document entries proved important in the make-up of the catalogue, and together with the digital data, dominated the content. Entries representing published documents currently total around 32% of all entries and unpublished documents around 19%. In general, document datasets tended to be recent (printed less than 10 years ago), indicating a spate of production in the nineties.

Digital data (including GIS and time-dependent climatic data)

Digital data were divided into two format types, i.e. GIS and other, the two having different field structures. Time series was accommodated adequately in the "other digital data" format. A current approximate total for GIS entry is 9% and for other digital (including time series) 35% of all entries. The bulk of the digital metadata

records were originally contained not in the main section of the catalogue but in an appendix which followed the layout of DWAF databases and database catalogues (for all datasets for which they were custodians). However, the latest version of the catalogue has now included these as normal records. Many of the digital datasets were collected or started more than 10 years ago, although ongoing time series datasets obviously cover the period till the present.

Simulation models

Models make up 28 entries (2% of all entries). They proved to be a category that was more difficult to collect than others, perhaps because relatively few people deal with these, and those who do consider them unfinished (until thoroughly tested). These entries covered both old and recent models.

Collections (including museum collections, videos, and photographs)

Collection entries (37) comprised a small percentage of the total (3%). They included metadata for hydrology (photographs, slides, video) (6 entries); water quality (satellite photos) (single entry); riparian vegetation (photographs, slides, videos) (5 entries); invertebrates (photographs, slides; preserved animals) (3 entries); fish (photographs, slides; preserved animals) (5 entries); legal/decision support systems (photographs, slides, checklists, videos) (10 entries).

Added to the four main categories of information were abstracts of key documents and fixed- point digital photographs from the KNP. The latter constitutes actual data and would normally be outside of the scope of an initiative such as this. However, the need for their usage was real and the chance of them being digitised without the help of the programme was small.

Abstracts

There was significant demand for abstracts of the catalogue-listed documents. Copyright holders were therefore requested to release their abstracts. Where abstracts were not available, interpretive abstracts were compiled by a collator from this project. About 32 documents were abstracted by the collator and 44 abstracts were from the copyright holders, these 76 comprising almost all of the "priority one" documents referred to below.

Abstracting was a time-consuming task. Although there are no copyright restrictions on creating abstracts, it was recognised that authors or publishers could be displeased with the result. Where possible the abstracts, written by the collator, were referred to the original authors for comment. In some cases this encouraged the authors or publishers to release their own abstracts to the KNPRRP. For those documents where abstracts were not available, prioritisation for abstracting by the collator was done using importance ratings in a KNPRRP status review by O'Keeffe and Coetzee (1996). Only the documents rated as "priority one" were abstracted due to time and budget restrictions.

Fixed-point digital photographs

The KNP is in possession of a 15-year twice-yearly series of fixed-point photographs, including views of rivers and riparian vegetation. This key dataset is very tedious to access, in the photonegative form. In addition, original negatives and prints are not allowed out of the immediate archival environment. The KNPRRP agreed to digitise these and index these for easy time-series

reference of the same locality. As a result, the fixed-point photo series was captured under relatively standardised digitising conditions, and stored in .jpg format on CD-ROM. Parties interested in the use of this data can contact Dr. Holger Eckhardt at holgere@parks-sa.co.za for further details.

Transferable experiences

The promoters and sponsors of the KNPRRP were using the ISDM subprogramme as a pilot project with the express aim of developing processes and experiences which could be transferable to the management of other rivers in Southern Africa. The key experiences from the ISDM project are thus outlined below.

Personal visitation

The decision to pay personal visits to all information custodians proved cost-effective. In almost every case the desired metadata were acquired. Past experience in similar ventures indicated that indirect methods, no matter how well-structured, were unlikely to yield much more than half to two-thirds of the material sought. Another very valuable feature of the personal-interview method is the goodwill and enthusiasm generated. The collators made a point of being good ambassadors for the programme. The trust engendered in this way appeared to place the ISDM subprogramme and indeed the whole KNPRRP in a position where custodians still continue to collaborate constructively, two years after the completion of the project. It is believed that the decision to engage in personal visitation has been a major contributor to this positive state of affairs.

Intellectual property and data sharing

Fewer problems than anticipated were experienced in the intellectual property realm, partly because of explicit contracts between funders and researchers. Significantly, there were also discernible shifts in opinion during Phase II, brought about by positive attitudes towards co-operation and data-sharing in the programme. The collator working inside the study area "unearthed" several unexpected datasets and several new custodians. It is believed that two of the major factors which lead to information staying "undetected" were the previous political fragmentation of the area and the widespread and obvious lack of commitment to any form of integrated catchment management right up until the early, and in many cases mid-nineties. Strong protectionist attitudes to data ownership prevailed at the time in South Africa. Data holders appeared to believe that it is wiser, more useful and more influential to keep their data to themselves than to share it. The authors believe it unlikely that such "unearthed" data will become "reburied" in the medium-term future, because of changes in attitudes, regional networking, and ongoing metadata cataloguing. Also, the availability of the catalogue enabled "outsiders" to the programme to gain a useful picture of the scope of the interdisciplinarity inside the programme.

Feedback to contributors

The importance of feedback to contributors of information cannot be over-emphasised. Once it was estimated that about 80% of all data had been catalogued, a hard-copy catalogue was prepared and circularised to all stakeholders as an interim product. This was accompanied by a database version, a search engine and an evaluation questionnaire. Approximately 50% of the questionnaires

were returned. The returns generally reflected positively on the necessity for, and quality of, the products. After this distribution exercise the ISDM subprogramme continued to trace and describe the estimated remaining 20% of datasets, and put much effort into devising and implementing methods to keep the catalogue up to date. These efforts benefited from the goodwill created by the progress evident and by the opportunity afforded by ISDM's feedback to contributors.

Training of collators

The fact that collators needed to be trained in various skills (e.g. database management, abstracting) and picked up interpersonal skills during the collation process, proved to be of value not only to the programme but also for their own personal development and development in other such programmes. This will almost always be the case when a programme needs to start off with willing but relatively inexperienced collators.

Use of the metadata within the KNPRRP

Through the production of the metadata catalogue ISDM forged links with all individuals and organisations who had previously researched water-related issues in the KNP rivers. From the questionnaire survey it was clear that many of them were using the catalogue for their purposes. In addition the subprogrammes within the KNPRRP made significant use of the catalogue. ISDM worked closely with the KNPRRP Secretariat which was responsible for maintaining project lists and the associated administration. Linkages between ISDM and this Secretariat resulted in setting up link keys so that all datasets related to a KNPRRP project could be traced.

The DSSDM (Jewitt and Görgens, 1995) made use of the metadata catalogue for their purposes and also provided valuable guidance to assist populating the model-related metadata. RDM made extensive use of the catalogue during their research for a status report on the KNPRRP (O'Keeffe and Coetzee, 1996). Rather than attempting all tasks necessary for comprehensive achievement of the set goals, the KNPRRP has concentrated on key projects, synthesis of existing information and the development of a coherent research management framework. Due to this as well as the fact that augmentation of internal activities relied on research and monitoring funded and motivated outside the programme, certain strategic information gaps were identified by this report, which also outlined priorities for further work.

The senior author of the previous status report (Görgens, 1998) for an earlier phase of the KNPRRP (Görgens and Lee, 1992) was of the opinion that the coverage of O'Keeffe and Coetzee was able to be far more comprehensive because almost all available sources of information could be traced as a result of metadata coverage. A further linkage between RDM and ISDM was that researchers new to the programme visited ISDM for an introduction to the metadata initiative, and were able to assess what was available, and what was expected from them in this regard.

Researchers on the TITT subprogramme placed the metadata on the KNPRRP website at the CCWR. The metadata catalogue was also linked into the Integrated Catchment Information System (ICIS) developed and driven by the KNPRRP.

An agreement was made with the IT group based at the WRC (previously operating as the South African Water Information Centre), who manage the WATERLIT bibliographic database, to collaborate in terms of the sharing of data. Although initially cautious of the wide definition of data sources adopted by ISDM, the WRC was encouraged by DWAF to add references relating to all printed information released by the KNPRRP to WATERLIT. It was also agreed that references to non-text sources, e.g. specimen collections, photo galleries, etc. will also be included in the database.

Involvement of ISDM in external activities such as congresses and road shows played an important role both in dissemination of useful knowledge gained in our own prototyping, and in attracting helpful criticism. Thus, for instance, a poster was presented on planned and existing products of ISDM (Coetzee et al., 1996) at a subcontinental conference which took place halfway through Phase II. The overall "role and position in the team" of the ISDM subprogramme in the KNPRRP was also discussed at the same meeting (Biggs et al., 1996).

Conclusions and recommendations

This effort served the immediate objectives of the KNPRRP's Phase II and also a wider community of researchers and stakeholders. Given that much development has taken place in the information industry since the start of KNPRRP Phase II in 1994, the authors would like to highlight the following concluding overall recommendations in a changing world.

Metadata collection across a wide range of data types would still be seen as the most affordable and efficient strategy to allow everyone in such a programme to access all available material. Recent developments in networking and mass storage technology allowed the updating functionality of this project, as well as any extensions beyond metadata, to be widely and easily accessible. Extensions of the metadata listings, allowing live connectivity with sites prepared to make raw or analysed data available, can also be built, given appropriate intellectual proprietorship arrangements.

Although the technological advances have been most fortunate, however, the key to the success of such a project still remains suitable and motivated people who form the core of the initiative. Enthusiasm on the part of data collators/information systems staff builds further enthusiasm amongst contributors and users. Staff who deal directly with metadata acquisition at a personal level should have the correct interpersonal skills and approach, and actions stimulating "buy-in" by participants should be initiated early in the process.

Acknowledgements

References

BIGGS HC, O'KEEFFE J and GÖRGENS A H M (1996) Information systems management in overall programme function - A case study from the Rivers Research Programme. *Proc. of the Vic Falls Conf. on Aquatic Systems and Int. Symp. on Exploring the Great Lakes of the World (GLOW) Food-web Dynamics, Health & Integrity*. 15-19 July, Elephant Hills Hotel, Victoria Falls, Zimbabwe.

COETZEE Y, LEFOTHLA W, UYS M, FREITAG S, VAN DER MERWE M and BIGGS H C (1996) Poster: Existing and planned products of the Information Systems Subprogramme (of the Kruger National Park Rivers Research Programme). *Proc. of the Vic Falls Conf. on Aquatic Systems and Int. Symp. on Exploring the Great Lakes of the World (GLOW) Food-web Dynamics, Health & Integrity*. 15-19 July, Elephant Hills Hotel, Victoria Falls, Zimbabwe.

JEWITT GPW and GÖRGENS AHM (1995) Integration of multi-disciplinary predictive tools for management of the rivers of the Kruger National Park - South Africa. *Proc. of the 7th S. Afr. Natl. Hydrol. Symp., SANCIAHS*, Rhodes Univ., Grahamstown, September.

GÖRGENS AHM (1998) Personal communication. Department of Civil Eng., Univ. of Stellenbosch.

GÖRGENS AHM and LEE J (1992) Kruger National Park Rivers Research Programme: Situation Statement and Management Assessment. Ninham Shand Inc. Report. 157 pp.

O'KEEFFE JH and COETZEE Y (1996) Status Report on the Kruger National Park Rivers Research Programme: A Synthesis of Results and Assessment of Progress. Water Research Commission Report No. 711/1/96. 63 pp.

Available on website **http://www.wrc.org.za**