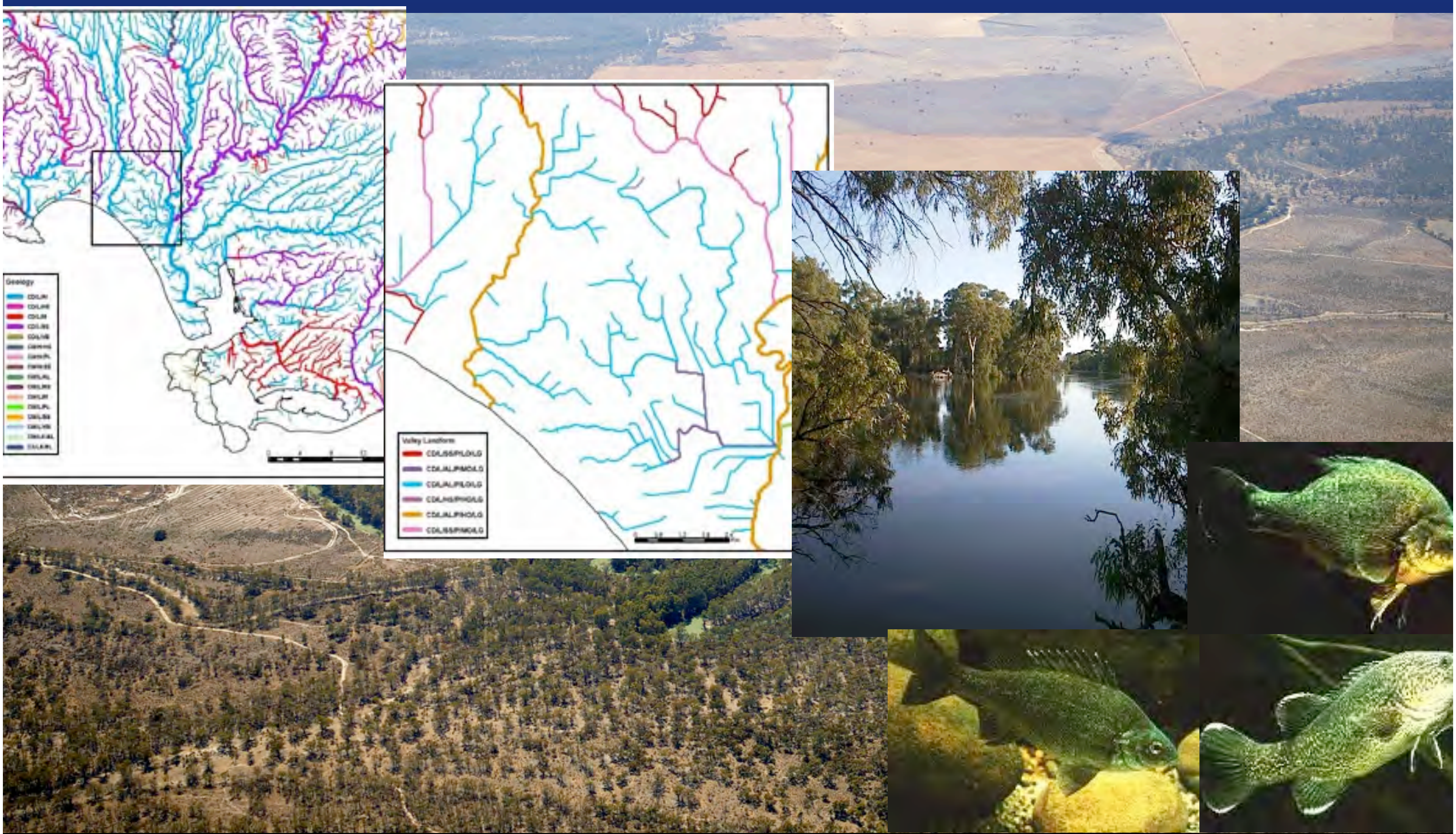
DEPARTMENT OF
SUSTAINABILITY AND ENVIRONMENT



MD740:

Spatial Management Model of Murray-Darling Fish Populations

Report to the Murray-Darling Basin Commission,
Native Fish Management Strategy

Belinda Cant, Jane Elith, John Leathwick, Yung En Chee, Julia Taranto,
Mark Case and Janet Pitchard

Arthur Rylah Institute for Environmental Research
July 2007

MURRAY-DARLING BASIN COMMISSION

Victoria The Place To Be

THE UNIVERSITY OF MELBOURNE

NIWA Science

aquatic habitat rehabilitation making more fish ...naturally

NSW DEPARTMENT OF PRIMARY INDUSTRIES

ISBN **978-1-74208-269-1**

Acknowledgements

We are very grateful for the input of the following people:

Attendees at the workshop:
NSW
Cassie Burns (Wetland Care Australia)
Paula D'Santos (Catchment Management Authority)
Ivor Growns (Department of Natural Resources and Environment)
Eren Turak (Environmental Protection Authority)

VIC
Wayne Tennant (or representative)
Peter Griffeon (Consultant to Arthur Rylah Institute, Department of Sustainability and Environment)
Matt White (Arthur Rylah Institute, Department of Sustainability and Environment)
Charles Todd (Arthur Rylah Institute, Department of Sustainability and Environment)
Dave Ramsey (Arthur Rylah Institute, Department of Sustainability and Environment)

SA
Annette Doonan (South Australian Research & Development Institute)
Malcolm Knight (South Australian Research & Development Institute)
Qifeng Ye (South Australian Research & Development Institute)
Angelo Tsolos (South Australian Research & Development Institute)
Sunil Sharma (South Australian Research & Development Institute)

QLD
Simon Linke (University of Queensland)
Dr John Kirkwood (Department of Primary Industries)
Sanjeev Kumar Srivastava (University of the Sunshine Coast)

Project members:
Dr Janet Pritchard (Department of Primary Industries),
Mark Case (Department of Primary Industries)
Dr Michael Smith (Arthur Rylah Institute, Department of Sustainability and Environment)
John Koehn (Arthur Rylah Institute, Department of Sustainability and Environment)
Dr. Graeme Newell (Arthur Rylah Institute, Department of Sustainability and Environment),
Dr John. R. Leathwick (NIWA New Zealand)
Dr Jane Elith (University of Melbourne)
Dr Yung En Chee (University of Melbourne)
Belinda Cant (Arthur Rylah Institute, Department of Sustainability and Environment)
Fern Hames (workshop facilitator) (Department of Sustainability and Environment)

Steering Committee:
John Harris (MDBC)
Peter Jackson (MDBC)
Richard Allan (MDBC)
Mark Lintermans (MDBC)

Funding agency:
Murray-Darling Basin Commission

Contents

List of figures

List of tables

Executive Summary

The Murray-Darling Basin Commission's (MDBC) Native Fish Management Strategy (NFS) aims to increase the abundance and diversity of native fish across the Basin to 60% of pre-European levels in the next 50 years. Six driving actions[1] are used to facilitate these increases in abundance and diversity. To maximise the cost effectiveness of these driving actions, resources (funding) need to be allocated to optimise the response of fish populations or communities to management actions. There are two main requirements for optimal resource allocation: 1. Applied knowledge of factors affecting fish populations and communities, and 2. The ability to select areas characterised by these factors for rehabilitation activities. One or both of these requirements form the basis of many native fish research, monitoring and rehabilitation projects undertaken by the MDBC and other agencies and research groups across the Basin. In order to add value to these projects the NFS envisaged developing a spatial database. The aims of this database or information system are:

i). to collate data relevant to fish management from across the Basin to aid decision making and data sharing among agencies and research groups

ii). to provide the capacity for dealing with new data (from more locations and for other taxonomic groups, physical processes and so on) into the long term future, including developing strong linkages to other MDBC initiatives such as the Sustainable Rivers Audit.

iii). to provide a structure and data capacity that allows interrogation and analysis of spatial data at a number of scales across time and space.

iv). to provide outputs that meet the needs of the range of likely users and applications. These would include, for example, (a) river managers in catchment management authorities requiring qualitative data on locations of important communities and processes, and (b) experts requiring quantitative analyses of information for complex or broader policy applications.

An important part of the process of designing a database is to set it up so it suits the data and the system in question, and can provide information in a useful format for likely end uses. The most straightforward use, summarising data and making maps, can be achieved relatively easily. However, recent advances in both geographic information systems (GIS) technology and in analytical modelling methods allow more complex but still very useful and

[1] Driving actions are: 1. Rehabilitating fish habitat; 2. Protecting fish habitat; 3.Managing riverine structures; 4. Controlling alien fish species; 5. Protecting threatened native fish species; 6. Managing fish translocation and stocking

achievable outcomes. For example, knowledge about the ecological requirements of fish species can be modelled so it is spatially explicit over extensive areas. This could, for example, aid identification of optimal river reaches for rehabilitation activities and other interventions within the Murray-Darling Basin. The main advantage of modelled data over point records or polygons based on relatively few observations is that data (including physical and biological variables) can be extrapolated to every segment of stream (river reach) throughout the Basin. It can be done with strong ecological foundations, making the information more likely to be biologically meaningful compared with simpler geographic extrapolations of known records. With modelled information, an environmental classification and/or a species distribution map can indicate the state of each segment of the river system and which sites share those characteristics that may optimise the response of fish populations to rehabilitation activities. This broadens capacities to identify sites suitable for the allocation of resources, and to select rivers, subcatchments, species or communities that may benefit from non site-based regulatory intervention. In the course of identifying these sites, relationships between fish populations and factors affecting them are made explicit. In this way, an important by-product of database and model development is an increase in knowledge about the ecology of Murray-Darling Basin fish communities. It is the view of the authors that this outcome is a highly desirable endpoint for a freshwater information system.

The most significant element of GIS technology capable of properly representing rivers and their environments is a description of the stream network that then allows a set of characteristics to be linked to a stream segment. Most importantly, the impacts of those characteristics are linked both up- and down-stream (characteristics are said to accumulate through the stream network). Before stream networks were developed, many GIS analyses of characteristics such as water quality or fish presence assumed that a stream segment functioned only in relation to adjacent land, often without incorporation of slope. This lacked a proper definition of catchment area. In conjunction with other Basin-wide environment variables, the stream network and associated data sets will form a framework upon which other smaller scale datasets may be added.

Both within Australia and internationally there has been wide and varied development of GIS resources, but many are not well suited to river networks. This scoping study describes approaches to developing and using a suite of GIS tools to aid native fish management that have been successfully applied to decision making. We recommend developing a flexible, dynamic and integrated spatial information system that provides information in suitable formats for further modelling of biological and physical variables. A similar approach has been successfully applied to government decision making problems in other jurisdictions. These applications included modelled physical data, distributions of fish and invertebrates, and environmental classifications based on powerful modelling methods and conservation planning software. Regardless of the type of data now available for the Murray-Darling Basin, an information base structured in an optimal way can gradually expand to meet future data additions and modelling initiatives.

We recommend that, as data are added to a new freshwater information base, their quality should be analysed and recorded, because the quality of models (e.g. of the relationship between fish presence and a set of site-based characteristics or predictor variables) reflect the quality of the underlying data. Substandard data may adversely impact the quality and outcomes of management decisions. A user should be able to choose to reject some records if they do not meet the quality specifications for a given end-use. Data that would allow informed analysis of records in a spatial database are described, and a selection of potentially useful data sets is presented.

Three case studies highlight the value of an information system based on a stream network in building knowledge while using spatial data to aid decision making. While both rule-based approaches and quantitative approaches may be useful for solving different problems, the most important aspect of the database is that all the data are linked and accumulate through an accurately defined stream network. It is recommended that the information system be designed expressly to be used at a number of scales. Flexibility is the most important characteristic of the database, so that it can be used for a variety of purposes by a variety of users, and so that it can be integrated with a number of other institutions' and agencies' databases. This flexibility will also allow additional data to be added as knowledge of the ecosystems is increased. While developing an information system such as this would be highly costly if starting from scratch, the main element of this system, the stream network, has already been developed methodologically and is currently being refined so it is useful at appropriate resolutions. The fact that the methodology is firmly established and has progressed well so far creates a strong foundation for system development.

We recommend pursuing the information system approach as a tool for identifying sites for optimal resource allocation. This will involve developing a database incorporating known quality biological and environmental data all linked to the stream network. These data could then be interpolated through the network and a classification developed that accurately discriminates between fish communities. Highly diverse sites, sites likely to have highly valued fish communities, or sites in which particular threats exist, could be identified, leading to resource allocation at appropriate sites. Another recommended element of the information system is fish habitat models that, depending on the availability of suitable data, would predict the likelihood of presence of particular fish species at each stream segment. This will refine identification of sites if particular species require targetting.

Glossary

Accuracy

The degree to which information on a map or in a digital database matches true or accepted values. Accuracy pertains to the quality of data and the number of errors contained in a dataset or map. In discussing a GIS database, it is possible to consider horizontal and vertical accuracy with respect to geographic position, as well as attribute, conceptual, and logical accuracy. The effect of inaccuracy and error on a GIS solution is the subject of sensitivity analysis. Accuracy, or error, is distinguished from precision, which concerns the level of measurement or detail of data in a database.

Architecture

An abstract technical description of a system or collection of systems. Modern software architectures employ interoperability interfaces to enable enterprises and whole industries to establish coherent, flexible, integrated information flows that can be implemented with heterogeneous but intercommunicating software systems. The OpenGIS Specification defines the interoperability interfaces that make it possible to include geographic information in these information flows. Conceptually based, architecture does not contain the level of detail needed for construction.

Attribute data

Descriptive information about features or elements of a database. For a database feature like census tract, attributes might include many demographic facts including total population, average income, and age. In statistical parlance, an attribute is a \`variable\` whereas the database feature represents an \`observation\` of the variable.

Base maps, data, or layers

Spatial data sets that provide the background upon which more specific thematic data are overlaid and analyzed. As inputs into a GIS, the term base map is usually applied to those sources of information about relatively permanent features including topography, soil data, geology, cadastral divisions, and political divisions. Within a GIS database, such information may become part of a land base to which other information is indexed and referenced.

Cartesian coordinates

Coordinates that differ from latitude-longitude coordinates in that the latter comprise a spherical (rather than planar) reference system.

Computer architecture

The functional composition of a system and its components, the interfaces between components, and interfaces with the external environment, including users and other systems.

Connectivity

A topological property relating to how geographical features are attached to one another functionally, spatially, or logically. In a water distribution system, connectivity would refer to the way pipes, valves, and reservoirs are attached, implying that water could be \`traced\` from its source in the network, from connection to connection, to any given final point. Functional, spatial, and logical connectivity are examples of relationships that can be represented and analyzed in a GIS database.

Coordinate system

Composed of a set of coordinate axes with a known metric. The concept 'metric of a coordinate space' consists of the set of mathematical rules that defines the relationships between the coordinate values and the invariant spatial quantities between points; for example, the mathematical rules (formulae) required for calculating angles and distances between points from coordinate values and vice versa.

Coordinate transformation

A mathematical operation on coordinates that includes a change of datum. The parameters of a coordinate transformation are empirically derived from a dataset containing the coordinates of a series of points in both coordinate reference systems. This computational process is usually &"over determined", allowing derivation of error (or accuracy) estimates for the transformation. Also, the stochastic nature of the parameters may result in multiple (different) instantiations of the same coordinate transformation.

Coordinates

A tuple of ordered scalar values that define the position of a single point feature in a coordinate reference system. The tuple is composed of one, two or three 'ordinates'. The ordinates must be mutually independent and their number must be equal to the dimension of the coordinate space; for example, a tuple of coordinates may not contain two heights.

Coverage

A feature that associates positions within a bounded space (its spatiotemporal domain) to feature attribute values (its range). GIS coverages (including the special case of Earth images) are two- (and sometimes higher-) dimensional metaphors for phenomena found on or near a portion of the Earth's surface. A coverage can consist of a set of features or Feature Collections. Earth images are seen as Grid Coverages that contain features whose geometries are of type “set of cells”; or “set of pixels”; (surfaces).

Coverage model

The basic model for how earth information may be represented as raster or grid coverages (e.g., an image or digital terrain model).

Database

One or more structured sets of persistent data, managed and stored as a unit and generally associated with software to update and query the data. A simple database might be a single file with many records, each of which references the same set of fields. A GIS database includes data about the spatial locations and shapes of geographic features recorded as points, lines, areas, pixels, grid cells, or TINs, as well as their attributes.

Data mining

The extraction of hidden predictive information from large databases for the purposes of knowledge discovery and predictive modelling. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g. rule based systems) and opaque in others such as neural networks.

Dataset

Any collection of related data, usually grouped or stored together.

Error

In a GIS database, a spatial or attribute value that differs from the true value. Error may also be understood as the totality of wrong or unreliable information in a database. Spatial errors are mainly errors in position (feature coordinates are wrong) and topology (features do not properly connect, intersect, or adjoin). Attribute errors are wrong quantities or descriptions associated with features, or missing or invalid values. Errors enter a GIS database through various processes, including data collection (for instance, flawed instruments); data conversion (for example, map digitizing mistakes); data entry and editing; data integration (for example, mixing data at different scales); spatial data processing (for example, inaccuracies caused by generalization); and data analysis (for example, features assigned to inappropriate categories on the basis of flawed criteria).

Feature layer

[data analysis] A layer that references a set of feature data. Feature data represents geographic entities as points, lines, and polygons

Flow regionalization parameter

A flow parameter such as the minima of a flow duration curve may be regionalized by using morphoclimatic characteristics of the drainage basin. The flow duration characteristics major flow gauging stations are first parameterized. Using statistical techniques, the geographic variation of each parameter of the best fitted flow duration model may be explained in terms of a set of variables readily measured in the absence of gauging stations such as mean annual areal precipitation, the drainage area, the hyposometric fall and the length of the main river course from the divide of the basin to the site of interest. The regionalized functions are used to synthesize flow duration curves at other locations within the hydrologically homogeneous region. (Mimikou and Kaemaki 1985).

Geodatabase

One or more structured sets of persistent data, managed and stored as a unit and generally associated with software to update and query the data. A simple database might be a single file with many records, each of which references the same set of fields. A GIS database includes data about the spatial locations and shapes of geographic features recorded as points, lines, areas, pixels, grid cells, or TINs, as well as their attributes.

Georeference

Description of a location relative to the Earth usually provided as coordinates.

Grid

[ESRI software] An ESRI data format for storing raster data that defines geographic space as an array of equally sized square cells arranged in rows and columns. Each cell stores a numeric value that represents a geographic attribute (such as elevation) for that unit of space. When the grid is drawn as a map, cells are assigned colours according to their numeric values. Each grid cell is referenced by its x,y coordinate location (c.f. feature layer).

Intergratability

is the ability to integrate data across multiple themes to improve its usability. Many projects require several datasets to be integrated and overlain (GIS) in order to show relationships. Use of a common datum, projection and data model components allows for such integration to occur

Layer

See Also : annotation layer, CAD layer,
feature layer, network analysis layer, network layer, raster layer, TIN layer

[data structures] The visual representation of a geographic dataset in any digital map environment. Conceptually, a layer is a slice or stratum of the geographic reality in a particular area, and is more or less equivalent to a legend item on a paper map. On a road map, for example, roads, national parks, political boundaries, and rivers might be considered different layers.

[ESRI software] In ArcGIS, a reference to a data source, such as a shapefile, coverage, geodatabase feature class, or raster, that defines how the data should be symbolized on a map. Layers can also define additional properties, such as which features from the data source are included. Layers can be stored in map documents (.mxd) or saved individually as layer files (.lyr). Layers are conceptually similar to themes in ArcView 3.x.

Metadata

Information that describes the content, quality, condition, origin, and other characteristics of data or other pieces of information. Metadata for spatial data may describe and document its subject matter; how, when, where, and by whom the data was collected; availability and distribution information; its projection, scale, resolution, and accuracy; and its reliability with regard to some standard. Metadata consists of properties and documentation. Properties are derived from the data source (for example, the coordinate system and projection of the data), while documentation is entered by a person (for example, keywords used to describe the data).

Object

1. [data models] In GIS, a digital representation of a spatial or nonspatial entity. Objects usually belong to a class of objects with common attribute values and behaviours.
2. [programming] In object-oriented programming, an instance of the data structure and behaviour defined by a class.
3. [software] In computing, a piece of software that performs a specific task and is controlled by another piece of software, called a client. For example, an object is often the interface by which an application program accesses an operating system and other services.
4. [ESRI software] In ArcMap, ArcScene, or ArcGlobe, the camera, view, table or layer to which an animation track is attached.

Pfafstetter coding system

An hierarchal characterisation of streams where watersheds are delineated from junctions on a river network. Level 1 watersheds correspond to continental scale watersheds. Higher levels (levels 2, 3, 4, etc.) represent ever-finer tessellations of the land surface into smaller watersheds, which are sub-watersheds of lower level watersheds. Each watershed is assigned a specific Pfafstetter Code based on its location within the overall drainage system and on the total drainage area upstream of the watershed’s outlet.

According to the Pfafstetter system, watersheds are divided into 3 types - basins, interbasins, and internal basins. A Pfafstetter basin is an area that does not receive drainage from any other drainage area; a basin contains the headwater of the river reach for which the watershed is defined. Conversely, a Pfafstetter interbasin is a watershed that receives flow from upstream watersheds. Finally, an internal basin is a drainage area that does not contribute flow to another watershed or to a waterbody (such as an ocean or lake).

Predictor variable

The variable that is manipulated by the experimenter. By attempting to isolate all other factors, one can determine the influence of the independent variable on the dependent variable

Physiographic

The physical features of the land, in particular its slope and elevation.

Snap

Spatial adjustment method found in ArcMap that aligns features along the edge of one layer to features along an adjoining layer. So all features along shared borders of layers have the same edge locations.

Spatial Data Infrastructure

The term “spatial data infrastructure” is often used to denote the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data. A spatial data infrastructure provides a basis for spatial data discovery, evaluation, download and application for users and providers within all levels of government, the commercial sector, the non-profit sector, academia and the general public.

The word infrastructure is used to promote the concept of a reliable, supporting environment, analogous to a road or telecommunications network. Spatial data infrastructures facilitate access to geographically-related information using a minimum set of standard practices, protocols, and specifications. Spatial data infrastructures are commonly delivered electronically via the internet.

Spatial reference system.

Position on or near the Earth's surface can be described by spatial reference systems. These are of two basic types: those using coordinates; and those based on geographic identifiers (for example postal addresses, administrative areas).

Topology

The arrangement for how point, line, and polygon features share geometry. Topology is employed in order to:

- Constrain how features share geometry. For example, adjacent polygons such as parcels have shared edges; street centerlines and census blocks share geometry; adjacent soil polygons share edges; etc.
- Define and enforce data integrity rules (e.g., no gaps should exist between polygons; there should be no overlapping features; and so on).
- Support topological relationship queries and navigation (e.g., to navigate feature adjacency and connectivity).
- Support sophisticated editing tools (tools that enforce the topological constraints of the data model).
- Construct features from unstructured geometry (e.g., to construct polygons from lines).

Sources for glossary:
http://www.opengeospatial.org/ogc/glossary/
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.gateway

Introduction

The project aims to scope the production of an information system which incorporates a suite of GIS tools and resources capable of assisting with native fish management decisions and other Murray-Darling Basin Commission tasks including

1. investment prioritisation,
2. identification of areas of high conservation value, and
3. areas likely to respond strongly to rehabilitation.

As the project progressed it became clear to the authors that the Murray-Darling Basin Commission (MDBC), and the Native Fish Management Strategy (NFS) within it, will benefit greatly from the development of an environment and philosophy of long term, integrated spatial data management. This environment will closely link decision making to empirical and modelled relationships between entities of concern (assets and threats) within the Basin and it will also be dynamic to accommodate both changing needs and new data collection.

An integrated spatial information system will contribute substantially to the NFS. Integration of existing data into GIS databases, and the ability to represent these data suitably for informing management issues, is an emerging priority for many natural resource management (NRM) authorities. Spatial models integrated into information systems offer an analytical environment than can sensibly combine a range of historical and current knowledge providing a useful tool for prioritisation and strategic planning.

In NSW, the Department of Natural Resources is developing a ‘Land Management Database’ that compiles on-ground works, activities, investments, scientific studies and assessments of condition into one integrated data library for each NRM region. By reviewing the assets, threats and investments across each region, NSW Catchment Management Authorities (CMAs) will soon be able to strategically plan their investments and prioritise areas for conservation or rehabilitation.

In Victoria, key data and statistical models of the extent and condition of native vegetation have been integrated in a spatial information system or geodatabase. A major project using these data seeks to optimise the protection and enhancement of vegetation on private land. This involves assessments of relationships between data objects (variables) in the database such as depletion, extent, condition and configuration of vegetation, and analyses of the effects on these of regulatory and economic tools such as government incentives. Recently, Habitat Hectare data (Parkes 2004), remote sensed data (e.g. Landsat) and other GIS data, were used to rescale the Habitat Hectares metric from the paddock to the regional scale. This study is currently being extended to give state-wide coverage. The completed datasets will provide a key source of information for NRM management across the State that can be used both for on-ground decision-making and as a baseline for estimating overall progress (i.e. as an input to "Net Gain" accounting).

Similar to these two terrestrial examples, spatial datasets or data layers have the potential to provide information about fish and the environmental and management factors that influence the viability of fish populations and the integrity of fish communities in the Murray-Darling Basin. The MDBC has already commissioned and managed several GIS projects to date. It is now an opportune time to move from a largely descriptive to a more predictive approach to data exploration and spatial analysis, taking advantage of recent advances in modelling and GIS methods. Useful patterns could be extracted from spatial datasets if relevant environmental (physical) data are developed, and the species data are appropriately handled.

The principal tasks for moving towards an integrated spatial information system were discussed in a workshop for this scoping study (for detailed notes see Appendix 2). These include:

- critically reviewing and prioritising the potential data layers relevant to fish in the Murray-Darling;
- evaluating the relative merits of a rigorous quantitative approach versus a more qualitative expert rule approach;
- refining a conceptual model to underpin and guide the management model;
- outlining the chronology of steps required to produce a functioning spatial data model;
- clearly identifying the range of management outputs that the model is likely to provide; and
- providing recommendations to refine the objectives, methods and outcomes of the project.

Specifically the workshop discussed the following which formed the basis for this report:

<u>1. Essential requirements of a spatial information system:</u>

- Must be hierarchical and able to operate at a range of scales (site, regional and catchment).
- Models used within it should be appropriately matched to end-user requirements – front end must be user-friendly and make the model outputs interpretable to the user, and be targeted towards real management needs.
- Any models should be used iteratively, not just to produce predictions and maps but to expose limitations in, and encourage refinement of, the data that is available for management.
- The accuracy of model outputs should be evaluated at the required spatial resolution.

2. Key challenges:

• Access to data across jurisdictions and institutions.
• Availability of expert modellers
• Making the models accessible to model-users so they can be understood and used appropriately.
• Immediate management needs versus building long-term capacity in the model.
• Data quality control – being able to quantify the accuracy of data (locational, attribute, resolution, etc).
• Uneven availability of fish sampling data across regions.
• Identifying the most useful set of environmental variables, and constructing these

3. Important points to keep in mind:

• Resource demands increase with the requirement for a higher level of scientific rigour.
• Robust geodatabase design is essential
• Data management systems need to be set up so that the data can be summarised in ways that are congruent with the behaviour of river systems.
• The storage of point observations of species occurrence is relatively straight forward, though all such observations need to be associated with their relevant river segment
• The representation of river and catchment topologies in a GIS is more complex, and needs to accommodate the network nature of river systems which is both linear and additive.
• The ability to cross-link each river segment both to its immediately contributing and downstream neighbours, as well as its immediately contributing and upstream catchment is also important

4. Identifying the advantages and disadvantages of different modelling approaches

• composite indices (weighted rules approach/ multi-criteria analysis) e.g. NSW TOOLS2;
•Environmental-based classifications (Snelder 2004);
• Environmental filters (Chessman 2006);
• Species distribution models (e.g. Leathwick et al. in press)

5. Data on fish populations, river network, hydrology, geomorphology and habitat, water chemistry of the Murray- Darling Basin;

• What data are available and its value?
• Prioritisation – which should be added to a database first?;
• Comprehensiveness (availability across the MDB);
• Internal consistency (any changes – e.g. in survey methodology);
• Currency (how old is the information available);
• Data gaps and integration issues;
• Ease/cost of collation versus creation.

6. Priorities for potential data:

• River data are stored in a network structure that allows linkage to

related upstream and downstream segments;
• River segments are able to be linked to their contributing catchments;
• All site data (biological, site descriptions etc) are georeferenced and snapped to river segments;
• Predictor variables are calculated for every river segment where possible, and can therefore be mapped;
• Some predictor variables are only measured at sites, but for many of these statistical models can be used to predict likely values throughout the river network;
• Biological data may include fish and other groups (e.g. invertebrates), and should have information on survey methods and effort etc.;
• Environmental predictors need to be
functionally relevant to the species / entity being modelled;
have good accuracy;
at scales relevant to processes in the system;
consistent across the whole network.

7. Characterising and coping with spatial and temporal errors and missing data.

8. Model output and how to evaluate model performance.

Whilst the original tenders for this scoping study outlined requirements for a “spatial management model”, further discussions clarified that "integrated information system" may be a better description of the key output of this project. This information system will contain both observed and modelled data, because models may be used to derive data layers. Models are useful because they can provide continuous data where only point data exists (Box 1) , and would be based on empirically derived relationships between objects. A model could be a simple prediction or classification of physical and environmental attributes, or it could be a prediction of the likely occurrence of fish species across areas of interest. The modelling software might be explicitly linked to the system, or may sit alongside of it, for use by expert modellers. Development of the integrated information system would require both setup of the database, entry of data, and modelling of required information.

The following report provides information on the essential characteristics of a freshwater information system for native fish with a particular focus on design of the database and datasets that may be useful (Chapter 2). This is followed by a description of three case studies illustrating the use of spatial information systems to aid decision making (Chapter 3). A recommended approach for phase 2 of the project is then provided (Chapter 4) and then a long term outlook is discussed (Chapter 5).

Box 1: Highlighting the importance of data for all river segments, compared with point data

The important features of environmental data for management and planning applications is that they are stored in a geographic information system that provides spatially referenced information across the whole region of interest. This means that any products derived from such data also supply information across the whole landscape, in contrast to point locations of site-based data that only inform users at the point of survey. The ability to plan across the landscape provides flexibility to consider all surrounding and upstream and downstream environments, to track changes, and set up ongoing sampling and monitoring. New data can be integrated easily because all locations in the landscape are represented in the environmental data, so new data can be linked to existing mapped information. It is sometimes suggested that point measurements of the environment can act as an alternative, because lands (or rivers) between sites can be assumed to be similar to the sampled sites, and potentially geographic interpolation techniques can be used to model this similarity.

Whilst this might work if sites were close together, the sometimes considerable distances separating sites introduces errors. Species data always exist as point locations, and need to be interpolated, because again, there are usually substantial distances between sites, and it is likely that the river sections between sampled sites do not provide constant habitat suitable for species observed at the sampled sites. Rather than use geographic interpolation methods, and because species primarily exist where they are because the habitat is suitable, results are much more robust if based on modelled the relationship between the species and its environment (see Chapter 3, case study 1, for an example). Such a model links species occurrences to functionally relevant environmental data, and uses this modelled link to provide "interpolated" estimates of the suitability of unsampled environments for the biota, as shown in Figure 3.4.

Data & design for the aquatic information system

Introduction

A well designed freshwater spatial information system (formerly spatial management model) for the NFS will initially be used to store, access, share and analyse data on fish and biological and environmental drivers of fish abundance and distribution. The freshwater spatial information system also has potential to be highly valuable as a basis for data storage and analysis for other initiatives and projects of the Murray-Darling Basin Commission (MDBC). The information system will directly support management of freshwaters and freshwater fish by providing evidence on which to base decisions which will help to optimise the allocation of resources.

In this chapter we explore data from several points of view: the data architecture that is most suited to river networks; typical properties of data that are usually available and the strengths and weaknesses of these for a range of outputs; and a report on the current status of data that might be usefully added to an MDB spatial information system.

Database architecture and institutional context

The development of the freshwater spatial information system was originally conceived primarily to draw together disparate datasets from across the Basin and to store data arising from NFS monitoring and research activities. However, it became evident from discussions at the workshop run for this scoping study that the design of the information system was the most important aspect to concentrate on.

The design of a database needs to be strongly coupled to the needs of the people using it and needs, above all things, to be flexible, so it can support a variety of users, the addition of new data at different scales, and new uses. Further, the information system needs to be designed to reflect the nature of the real system for which it stores data. Riverine ecosystems are best described as an hierarchical network with both upstream and downstream in-stream linkages and lateral linkages with associated catchments. The information system will store data in such a way that the level and position at which data sit in the hierarchical network and longitudinal and lateral linkages are explicit for each data point.

During the workshop, participants discussed the key characteristics of a freshwater information system that are essential for it to function well (Box 2)

Box 2. Essential characteristics of the MDBC freshwater information system

- An aim to develop a useful, overarching, multipurpose information system that evolves as data are acquired, user's needs change and methods of analyses are refined to meet those changing needs.
- The data structure based on a river network that is linked up- and down-stream, and to the surrounding catchment(s) and river basins within the MDB so that data can be summarized in ecologically meaningful ways. Data are to be linked in a hierarchical or nested fashion from the whole of Murray Darling Basin to individual river basins; catchments within river basins; subcatchments within catchments; and reaches within subcatchments.
- Flexible data architecture to allow additions of a wide variety of types of data at a number of scales and allows linkages to other state and federal agency's databases.
- An open or networked architecture that allows sharing and accessibility across the Commission, between jurisdictions and particularly to those people/agencies/institutions who provide data (this is the key to obtaining data).
- Quality control and metadata that includes derivation, accuracy, and extent required for uploading and caveats regarding certainty and purpose always linked to data and products that are downloaded.
- Institutional arrangements that ensure ongoing maintenance of the database and strong links between database designers, database managers, modellers and biologists and users to ensure relevant outputs.

Spatial data and spatial information systems for rivers – a general overview

Part of the intention of this scoping study was to assess practices of data collation and modelling both within Australia and internationally. This section gives that overview. The practice of collating spatial data and using it with or without models for management or planning has a longer history in terrestrial systems than freshwaters. and because much of the methodology is transferable, several examples here are terrestrial ones.

Typical characteristics of 'found' biological data

In many of the ecosystems found in the Murray-Darling Basin there is no long history of Basin- wide systematic, planned biological surveys, and the available data largely comprise:

1. planned surveys carried out for a particular purpose, usually over a small proportion of the full extent of the Basin. These generally use sampling methods suitable for recording species presence and absence, or abundance. Sometimes they are used as the basis for community mapping;

2. wide-ranging surveys without a clear survey plan, where all observed species (often those in one broad biological group – e.g. birds or vascular plants or fish) were recorded, often with poor locational precision, and in which absences of species from locations are generally not recorded systematically;

3. incidental ("presence-only") records, such as those recorded in museums or herbaria, where the presence of a species is noted and recorded – sampling of such records is generally non-systematic and uneven, often varies from species to species, depending on their collectability, and may include a large proportion of records that were collected because of their unusual nature e.g. they were not usually observed in that location;

Planned surveys are most likely to provide samples across geographic or environmental gradients, leading to the least biased view of the biota. Incidental records are often biased toward the most interesting species and the most accessible locations (including those closer to roads or tracks, and closer to towns). Typical data sets have a range of other problems including:

- not collated, and sometimes not yet recorded digitally;
- no information on survey method, survey effort, taxonomic accuracy, or location accuracy;
- collected over many years, meaning that the location may no longer be suitable for the species;
- species identifications not updated taxonomically;
- locations recorded in many different projections / datums, which may not be fully documented;
- sparse records in many parts of the geographic region of interest.

The lack of systematically collected data is a problem for conservation planning and management, because if the biological data are used directly to assess conservation value, then no such assessment can be made for non-sampled parts of the region. Alternatively, if models are used it is necessary (but still difficult) to deal with likely biases and errors in the data. Table 2.1 includes examples of studies that use data and try to deal with uncertainties and biases in the data.

Environmental data - requirements and typical properties

One aim in a spatial information system is to collate environmental data that are useful and ecologically relevant to the biota of interest. The concept of environmental filters is one useful approach for thinking about environmental data relevant to a species. This concept describes a regional or global species pool, a proportion of which is successively excluded by a set of environmental factors operating over a range of scales. A residual local assemblage that tolerates the set of local environmental factors characterising a site, are predicted to be present at that locality (Chessman 2006). In the case of fish, biogeographical barriers such as areas of high elevation that distinguish basins may exclude species from a site by preventing dispersal (Chessman 2006). Fish species also differ in physiological tolerances to temperature and salinity while differing habitat requirements lead to presence and absence being determined by discharge or presence of large woody debris.

There are many environmental factors correlated to fish species presence and abundance – for example, a list of variables that have proved powerful for explaining fish presence and absence in New Zealand are provided in the following chapter, in Table 3.1. These provide an example of the sort of variables that could be developed. In addition we explore in other parts of this chapter Australian data or software that are available and that may be useful. In a spatial information system relevant to management other spatial data relevant to tenure, disturbances and other factors will also be useful.

Many of the problems associated with 'found' biological data also apply to spatial environmental data. Some common shortcomings include:

- incomplete data that do not cover the whole landscape of interest
- lack of metadata describing what the data are, the purpose for which they were collected/developed, the method of collection/development, capture scale and accuracy, projection, the time period for which the data are relevant, status of the data (update frequency) and so on
- unknown, inconsistent and/or incompatible projections
- incomplete fields in attribute tables so that codes cannot be interpreted unambiguously
- defined river networks without topologically consistent characteristics such as directionality and connectivity (as is often the case with river networks digitised from topographic maps); defined stream networks in which segments/sections are not spatially linked to their local catchment areas
- site-based data at a small subsample of sites that are difficult to extrapolate to the whole region
- lack of environmental predictors that are functionally relevant to the biota of interest

Table 0-1 Examples of approaches to using data or models in conservation planning

Reference	Ecosystem/ species	Relevant features of the study
Bini et al (2006)	Anurans in Brazil	Explored the effect of sparse and incomplete sampling by comparing a reserve design based on currently known species with that based on additional data on inferred distributions of as-yet undiscovered species. Used habitat models and grid-based data.
Drielsma et al (2007):	Terrestrial, Australian but broadly applicable	Presents a method for combining habitat models with ideas of species movements to understand neighbour context and metapopulation capacities. Uses grid-based data.
Early and Thomas (2007)	Butterflies in Britain	Explores how to incorporate an understanding of population processes in conservation planning, so persistent populations could be identified. Used the multispecies prioritization tool ZONATION and weighted species according to the risks they faced and their international vs regional importance
Moilanen and Wintle (2006)	Terrestrial; fauna in NSW.	Investigates how to allow for uncertainty in reserve design, so that decisions are robust to the uncertainty in the underlying modelled species' distributions.
Scheuerell et al (2006)	Salmon in USA rivers	Use spatial information, a population model and proposed management scenarios to evaluate the likely population responses to habitat restoration.
Rondinini et al (2006)	Based on review and interpretation, not data	Explores the effect of different types of errors in data / species models on conservation planning outcomes, and recommends explicit reporting of likely errors to help decision-making.
Linke et al (2007)	Rivers of Victoria, Australia. Invertebrates	Used modelled invertebrate distributions to assess irreplaceability, condition and vulnerability of rivers and suggest priorities for reservation or restoration

Database design is a complex task that requires specialist skills. In addition, it is critical that database construction is carried out in collaboration with experts with experience in modelling and management. In particular, the likely required outputs from the data should govern the data architecture, which should be flexible enough to allow likely current and future uses. How these data are used in planning, whether via modelling or simpler summaries, is a separate question to how the data are collated and stored. While by no means definitive, Tables 2.2 and 2.3 highlight some of the desirable features of environmental and biological spatial information systems for freshwater systems.

At a federal level there is an awareness that database management requires dedicated resources and policies and procedures to ensure that data are treated as a valuable asset. Implementing such policies and procedures yields many benefits including ready access to reliable data for better decision-making, encouraging more extensive use of a valuable public resource, avoiding duplication of effort and maximising integration, interoperability and efficiency by adopting common standards for the collection and transfer of data, improving communications and facilitating partnerships (Commonwealth of Australia 2003).

Table 0-2 Desirable features of an environmental spatial information system for freshwater systems

Feature	Explanation of importance
River network to be derived from drainage analysis using the best-available hydrologically correct digital elevation model (DEM).	Enables delineation of river segments with topologically consistent characteristics such as directionality and upstream-downstream connectivity. This information enables the construction of an ordered river network.
Rivers should be represented in a network topology, where each river segment (defined as an individual river section between two river junctions) has a unique identifier and is linked to its immediate upstream and downstream neighbours with specific reference to their identifiers.	Need to be able to trace rivers upstream and downstream to properly account for relevant environmental and biological influences. For instance, to quantify upstream inputs to sites, identify barriers to fish migrations etc. River systems that are not represented as ordered networks are not able to express the essential features of rivers – that they are linear and additive in their connectivity.
Subcatchments and catchments should be identified in conjunction with the river network so that for any individual segment, its local subcatchment as well as total upstream catchment can be identified (see Figure 2-1).	The ability to link each river segment to its immediate subcatchment and upstream catchment is important for characterising catchment inputs to a site. For example, to enable areal inputs such as rainfall-runoff to be accumulated over the upstream catchment area of each network segment.
Environmental data for rivers should be stored at the scale of the individual river segment. This includes climatic, hydrologic, physiographic, geologic and land cover data etc. Segment-scale and catchment-scale environmental variables should be calculated for every river segment where possible. Relevant upstream and downstream attributes (e.g. distance to coast, maximum downstream slope etc.) should also be computed for every segment where possible.	This is the scale most useful for landscape-scale modelling and analysis of environments and species-environment relationships. As the data are associated with every river segment, analyses can be extended to the entire river system.
Other spatially distributed data that allow assessment of pressures on the river system and/or estimates of condition. For example, sediment and nutrient loads (e.g. see DeRose *et al.* 2003), anthropogenic flow regime alterations, point sources of pollution, urban development, agricultural and forestry inputs (e.g. see Stein et al .2002), native vegetation loss in catchments / river edges, and introduced fish.	Enables scenarios of change to be assessed, or spatially explicit analyses of the potential effect of removal of pressures on freshwater-dependent biota.

Table 0-3 Desirable features of a biological spatial information system for freshwater systems

Species data – presence-absence data are preferable to presence-only data	Presence-absence data has greater information content than presence-only data. For example, when species use habitats proportionally to their suitability, absence data helps to enhance model predictive performance (Brotons *et al.* 2004).
At a minimum, species data should include the sampling date, sampling method and catch effort. Species data should be geo-referenced (projection stated) and/or have its location recorded in sufficient detail to enable it to be matched to a specific river segment in the river network. All site-based environmental data measured at the time of sampling (e.g. water temperature, pH, turbidity, concentration of dissolved oxygen, substrate type, quantity of woody debris etc.) should also be entered into the biological database.	Species data can then be linked to the river segment scale abiotic data stored in the environmental database and be used for modelling and analyses of species-environment relationships. As site-based data accumulate it may enable the development of statistical models capable of predicting the distribution of fine-scale environmental predictor variables (such as substrate) throughout the river network.

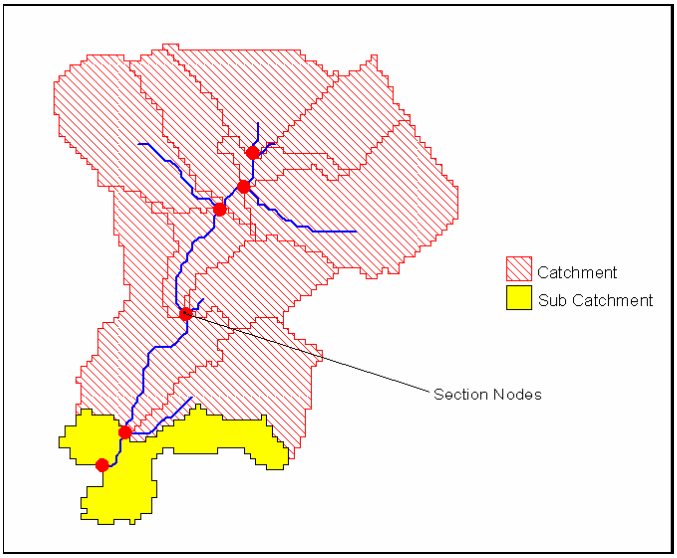


Figure 0-1 Example of the relationship between subcatchments and catchment
Source: Wild et al. 2005

Part of the utility of this data will be that a range of people can work on it and provide relevant research for the MDBC. With good infrastructure and an ability to provide high-quality data to interested users there comes capacity to interest new parties, to stimulate research questions and encourage research that will expand the knowledge base for decision-making by the MDBC. At the moment, because the data are so fragmented and there is no one secure, well-designed repository to the data, any researcher/natural resource manager faces too high an overhead to start collecting the information together, or what is done is done in isolation and may not be collected in a way that adds value to what already exists.

ANZLIC – the Spatial Information Council (http:/www.anzlic.org.au) is the peak intergovernmental organisation developing consistent policies on collection, management and use of spatial data in Australia and New Zealand. Their web-published policies and guidelines on 'best practice' in spatial data management provide clear and detailed protocols on geospatial standards, access and use. In particular, there is good advice on spatial metadata (which is information about data, describing the "what", "when", "who" and "how" of data). As a minimum requirement, data in should be accompanied by metadata conforming to the ANZLIC Metadata Guidelines (Core metadata elements for geographic data in Australia and New Zealand Version 2 – February 2001) with the data quality elements completed to a satisfactory level. The ANZLIC metadata guidelines may be broadly applicable to biological data, but additional elements may have to be included so that the methodological details associated with the collection/sampling and processing of the biological data are available to database users. The elements of New Zealand’s FBIS could serve as a useful template. These standards are important in order for end users to understand the data and its strengths and limitations.

One end-use of these data is likely to be modelling of fish distribution. Modelling is a specialised task, and needs experienced staff and/or collaborations with experienced researchers. There is wide variation in the performance and characteristics of different modelling methods and some are much more flexible and capable of expressing subtle ecological relationships than others (Elith *et al.* 2006). Using sophisticated modelling tools does not imply that the output needs to be incomprehensible to users; outputs relevant to the needs of the users can be supplied. See examples in the New Zealand case study (Chapter 3), in which sophisticated modelling methods were used, and these provided both important ecological insights and useful outputs for management. Outputs from modelling could include:

- models of species' distributions, that can then be used for understanding ecological relationships and for predicting the distribution of species across the river network.
- classification of species into community groups
- environmental classification of rivers, that have a hierarchical structure enabling use at widely varying spatial scales. Use of tools such as GDM allow classifications to be defined with reference to parallel biological data, maximising their ability to discriminate variation in biological patterns.
- prioritisation of catchments or sub-catchments for conservation management or restoration
- prediction of the likely outcomes of a range of management scenarios, including evaluation of which provide the most robust solutions given the uncertainties in the data and the system.

Data – priorities, availability and properties

Whilst it would be possible to include all available and relevant data in a spatial information system, it is prudent to prioritise data input. Initially the database should be populated with current high quality spatial data relevant to fish in the Murray-Darling Basin. These include biological data such as fish presence and absence data and other biological and environmental variables that influence the abundance and distribution of fish (see previous section and case studies in Chapter 3) or that are otherwise relevant to management. It is important to have data that can provide outputs that are accurate enough for required end uses. Whilst lower quality data might still prove useful, particularly if likely errors are recorded and taken into account in any analyses, it is sensible to put initial efforts into entering higher quality data.

Questions that could help assessment of the initial suitability of data are outlined in Box 3, and further important features are discussed in the previous section. Searching for data and assessing it is a large task. Unfortunately many data sets held in regional and research organizations in Australia are not fully documented and lack metadata. In these cases assessing in detail what data are held and the quality of that data would be an expensive and time-consuming undertaking. Preliminary information on available data is presented in the remainder of this chapter. It is the view of the project team that the resources required to find, collate, standardise and fix errors in State–regional- and research agency- based data sets will make the pursuit of this type of data costly (Mohammadi *et al*. 2006), with likely low returns of useful information for the investment. Nevertheless at some stage funding and time may be available for its assessment and inclusion. We suggest first focussing on data with known and useful attributes. In that way, the first products from the spatial information system would have a strong basis and their usefulness may demonstrate why data collation and cleaning at a regional level could be valuable. Other data that are capable of informing future questions could be added as time and need dictated. Because of the amount of data that could potentially be collated it may also be advisable to develop some stopping rules for searching for and collating datasets in later phases of the project.

The remainder of this chapter reviews biological and GIS data currently in existence. It is organised into: A. Data where quality is uncertain and further research is required to verify its value on a case-by-case basis, discussed as (i) publicly accessible data; and (ii) data that are not in the public domain and may be less easily accessed ; B. datasets that appear to be high quality and suitable for initial entry into the information system. The focus is on the data that are available and its quality, but not cost. As an overview, Table 2-4 presents a summary of broad themes of data that would be useful, and data that have been identified that appear useful.

Box 3: Relevant questions for assessing suitability of data for the database

A. For existing data:

- are the data georeferenced to the required standard? (this standard needs to be specified in relation to likely end uses);
- what is the spatial extent of the data?;
- are data are consistent across the whole stream network or can inconsistencies can be quantified?
- are suitable metadata available? (this could be developed into detailed specifications);
- do the data require extensive resources spent on further derivation or error correction?;
- - is access to the data unrestricted, or how is it restricted?
- - what is the cost of the data?;

B. For developing new data:

- can environmental variables be calculated for all river segments?;
- is environmental data:
  - functionally relevant to species of interest?
  - accurate at the scale at which they are to be used?
  - at scales relevant to processes in the system?