Report No. 2a **STANDARDS**

Data Management Standards

a. Internal File Systems

Preliminary report prepared by Arthur D. Chapman – May 2003.
Updated January 2004.

Background:

CRIA is beginning to accumulate Spatial data from a range of sources, as well as increasingly creating Spatial GIS layers in-house.

The present storage arrangements are ad-hoc with data stored in a range of locations, in a range of formats, and with varying levels of documentation. Often data is stored on the C-Drive of individual staff members, which creates a problem with back-up and access.

CRIA has also recently purchased ESRI's ArcView 8 for use by CRIA staff.

It would appear to be an opportune time for a set of internal data-storage and management standards to be adopted by staff in order for data to be generally available, be fully documented, and to be in a consistent format.

Options

A range of options exists for the storage of data. I have made a number of suggestions and recommendations below. These have been adapted from the Standard in use in Environment Australia and developed there by Neil Freeman in collaboration with staff. This proposed standard is open for discussion.

There are three levels of storage and documentation proposed: storage for final products that are "read only" in most cases, and data products that are still being worked on, but which may be regarded as permanent or semi-permanent, and scratch or temporary documents.

An Information Management Toolkit was recently developed for the Australian National Land and Water Resources Audit. This is a very detailed document, and one that has a lot of relevance to CRIA. I suggest that this report be examined and links made to it from the CRIA Website.

http://www.nlwra.gov.au/toolkit/contents.html

A. FINAL PRODUCT DATASETS

The intention of the storage system below is for **final product datasets**. It is not intended for working datasets – see part B for these.

It is recommended that once a data set is "final" it is placed into this structure and can only then be generally accessed as "read only".

In all cases – such final products must have full metadata documentation completed before being placed in the final data area. 100% of all data in this area is to be documented.

1. Location

It is recommended that all final data be stored in one location to allow general access to all staff. Ideally – a separate data drive would be set up for storage of final data products.

2. Directory structure

It is recommended that data be stored through directories and sub-directories following a series of data categories, themes, types, etc. These are explained below in a possible data structure.

Q:\[availability]\ [category] \ [theme] \ [type] \…

Not all directories may contain data: the filesystem standard is developed to be a broad scientific and infrastructural data classification to cover the broadest possible project requirements.

The themes, categories and types are explained below.

The "export" subdirectory, is for cases where a dataset is likely to be made available in compressed format for automatic download via the Web, or via ftp.

3. Examples of a possible structure

[For Satellite data]:

Q:\ sat_pub
\ **tm** (LANDSAT thematic mapper data)
\ **mss** (LANDSAT multi-spectral scanner data)
\ **noaa** (NOAA-AVHRR data incl. Derived data)

(each of these themes have a special subdirectory structure
best suited to each data product).

* due to the size of satellite images – special storage and image search routines may have to be developed if it is likely that a lot of data is likely to be acquired.

[For Packaged Topographic Datasets and Atlases]:

Q:\ top_pub
\ **topo250k** (topo250k data package)
\ **dcw1m** (digital chart of the world package)
\ **arcworld** (arcworld generalised data package)
\ **gaz** (gazetteer datasets)
\ **raster250k** (scanned paper 250k topographic maps – Geographicals)
\ **raster250k_utm** (scanned paper 250k topographic maps – UTM)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ top_pub \ gaz \ data \ gaz** (Gazetteer of Brazil: ARC/INFO coverage)

[For Climate Datasets]:

Q:\ cli_pub
\ **tersurf** (terrestrial climate surfaces)
\ **marsurf** (marine climate surfaces)
\ **clichang** (climate change data)
\ **stations** (weather stations and related networks)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ cli_pub \ stations \ data \ autows_net** (Automatic Weather Stations Network: ARC/INFO coverage)

[For Physical/Process Data]:

Q:\ phy_pub
\ **tergeosc** (terrestrial geoscience incl. Geology, geomorphology)
\ **margeosc** (marine geoscience incl. Benthic substrate)
\ **soils** (soils)
\ **hydrol** (hydrological data incl. Streams)
\ **freshwat** (fresh water data incl. Catchments)
\ **oceanphy** (oceanography - physical data incl. Wave height)
\ **oceanwq** (oceanography – water quality incl. Chemical concentrations)
\ **nathazrd** (natural hazards datasets incl. Firescars)
\ **landdeg** (land degradation datasets incl. Erosion and salinity)
\ **glaciol** (glaciological data)
\ **disturb** (disturbance datasets incl. River disturbance)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ phy_pub \ freshwat \ data \ brazdrain_250k** (Drainage Basins, Divisions and Regions: ARC/INFO coverage)

[for Biotic Data]:

Q:\ bio_pub
\ **landcov** (landcover datasets)
\ **veg** (vegetation datasets)
\ **florsite** (flora site datasets)
\ **faunsite** (fauna site datasets)
\ **marhabit** (marine habitat datasets)
\ **modflor** (Modelled predictions – flora)
\ **modfaun** (Modelled predictions – fauna)
\ **modsum** (Modelled summaries)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ bio_pub \ veg \ data \ pveg5m** (1:5m Present Vegetation of Brazil: ARC/INFO coverage)

[For Ecosystems Datasets]:

Q:\ eco_pub
\ **landsys** (landsystems datasets)
\ **marreg** (marine regionalisations)
\ **terreg** (terrestrial regionalisations)
\ **landscap** (landscape datasets incl. Wetlands)
\ **cstzone** (coastal zone data)
\ **mareco** (marine ecological datasets)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ eco_pub \ terreg \ data \ ibrb** (Interim Biogeographic Regionalisation of Brazil (IBRA): ARC/INFO coverage)

[For Terrain/Morphology Datasets]:

Q:\ ter_pub
\ **elev** (elevation datasets incl. DEM and contours)
\ **morphol** (derived products from DEM eg. slope and aspect)
\ **cstlines** (coastline datasets)
\ **bathym** (bathymetric datasets)
\ **cont9s** (50 and 100 metre contours – all of Brazil)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ ter_pub \ cstlines \ data \ cst_100k** (1:100,000 scale Coastline of Brazil: ARC/INFO coverage)

[For Socio-Ecconomic/Cultural/Administrative Datasets]:

Q:\ soc_pub
\ **infra** (infrastructure data incl. Roads and built-up areas)
\ **demogpop** (demographic and population datasets)
\ **culherit** (cultural heritage datasets)
\ **indherit** (indigenous heritage datasets)
\ **landuse** (landuse datasets)
\ **seause** (seause data incl. Shipping routes, channels etc)
\ **polwaste** (pollution and water management datasets)
\ **admgov** (Government administrative boundaries eg. State borders, LGA)
\ **admother** (other administrative boundaries)
\ **index** (data index – 100k, 250k, 1m datasets)
\ **tenure** (tenure boundaries incl. Reserves, RNEDB, World Heritage)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **export** (ARC/INFO compressed export files) [if required]
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ soc_pub \ tenure \ data \ wha_public** (World Heritage Areas: ARC/INFO coverage)

Filesystem Security:

Security is enforced on the Core Data Filesystems such that:
- Only members of the **gis** group can write data to them. This limits updating these filesystems to CRIA's Core Data staff.
- General CRIA staff and users can access the data in a read-only mode via [Atlas Biota/ArcView].

The security is implemented on the UNIX filesystems as:

World:Read Only, Group:Read+Write, Owner:Read+Write where Group=gis, Owner=[????]

B. INTERNAL WORKING DATASETS

The intention of the storage system below is for **incomplete and working documents**. It does not apply to scratch or temporary datasets, however, any dataset that is likely to be kept for more

than 6 months should be considered as other than "scratch" or "temporary" and should follow this standard.

It is recommended that once a data set is "final" it is fully documented and placed into the above structure.

Documents in this area should be documented as soon as possible, and it is recommended that 100% of all documents more than 6 months should be documented and 60% of all documents less than 6 months old should be fully documented.

1. Location

It is recommended that all working data be stored in one location to allow general access to staff. This also allows for regular backup, and allows for continuity of access when staff are away, or leave the organization.

2. Directory structure

It is recommended that data be stored through directories and sub-directories following a series of data categories, themes, types, etc. These are explained below in a possible data structure.

Q:\[availability]\ [category] \ [theme] \ [type] \…

Directories need not be set up for all categories if there is no data, however, it is recommended that the data structure follow that for the Final data products mentioned under "A." above.

The themes, categories and types are explained below.

As these are not final datasets – an "export" directory is not necessary. However, a "README" file should be included in all directories, providing basic information about the files in that directory. This is particularly important for data that is not yet fully documented with metadata, and provides users with information on access conditions etc. of the dataset. See below for an example of a standard README file.

3. Examples of a possible structure

[For Satellite data]:

Q:\ sat_int
\ **tm** (LANDSAT thematic mapper data)
\ **mss** (LANDSAT multi-spectral scanner data)
\ **noaa** (NOAA-AVHRR data incl. Derived data)

(each of these themes have a special subdirectory structure
best suited to each data product).

* due to the size of satellite images – special storage and image search routines may have to be developed if it is likely that a lot of data is likely to be acquired.

[For Packaged Topographic Datasets and Atlases]:

Q:\ top_ int
\ **topo250k** (topo250k data package)
\ **dcw1m** (digital chart of the world package)
\ **arcworld** (arcworld generalised data package)
\ **gaz** (gazetteer datasets)
\ **raster250k** (scanned paper 250k topographic maps – Geographicals)
\ **raster250k_utm** (scanned paper 250k topographic maps – UTM)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ top_ int \ gaz \ data \ gaz** (Gazetteer of Brazil: ARC/INFO coverage)

[For Climate Datasets]:

Q:\ cli_ int
\ **tersurf** (terrestrial climate surfaces)
\ **marsurf** (marine climate surfaces)
\ **clichang** (climate change data)
\ **stations** (weather stations and related networks)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ cli_ int \ stations \ data \ autows_net** (Automatic Weather Stations Network: ARC/INFO coverage)

[For Physical/Process Data]:

Q:\ phy_ int
\ **tergeosc** (terrestrial geoscience incl. Geology, geomorphology)
\ **margeosc** (marine geoscience incl. Benthic substrate)
\ **soils** (soils)
\ **hydrol** (hydrological data incl. Streams)
\ **freshwat** (fresh water data incl. Catchments)
\ **oceanphy** (oceanography - physical data incl. Wave height)
\ **oceanwq** (oceanography – water quality incl. Chemical concentrations)
\ **nathazrd** (natural hazards datasets incl. Firescars)
\ **landdeg** (land degradation datasets incl. Erosion and salinity)
\ **glaciol** (glaciological data)
\ **disturb** (disturbance datasets incl. River disturbance)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ phy_ int \ freshwat \ data \ brazdrain_250k** (Drainage Basins, Divisions and Regions: ARC/INFO coverage)

[for Biotic Data]:

Q:\ bio_int
\ **landcov** (landcover datasets)
\ **veg** (vegetation datasets)
\ **florsite** (flora site datasets)
\ **faunsite** (fauna site datasets)
\ **marhabit** (marine habitat datasets)
\ **modflor** (Modelled predictions – flora)
\ **modfaun** (Modelled predictions – fauna)
\ **modsum** (Modelled summaries)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ bio_ int \ veg \ data \ pveg5m** (1:5m Present Vegetation of Brazil: ARC/INFO coverage)

Q:\ eco_int
\ **landsys** (landsystems datasets)
\ **marreg** (marine regionalisations)
\ **terreg** (terrestrial regionalisations)
\ **landscap** (landscape datasets incl. Wetlands)
\ **cstzone** (coastal zone data)
\ **mareco** (marine ecological datasets)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ eco_ int \ terreg \ data \ ibrb** (Interim Biogeographic Regionalisation of Brazil (IBRA): ARC/INFO coverage)

[For Terrain/Morphology Datasets]:

Q:\ ter_ int
\ **elev** (elevation datasets incl. DEM and contours)
\ **morphol** (derived products from DEM eg. slope and aspect)
\ **cstlines** (coastline datasets)

\ **bathym** (bathymetric datasets)
\ **cont9s** (50 and 100 metre contours – all of Brazil)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ ter_ int \ cstlines \ data \ cst_100k** (1:100,000 scale Coastline of Brazil: ARC/INFO coverage)

[For Socio-Ecconomic/Cultural/Administrative Datasets]:

Q:\ soc_int
\ **infra** (infrastructure data incl. Roads and built-up areas)
\ **demogpop** (demographic and population datasets)
\ **culherit** (cultural heritage datasets)
\ **indherit** (indigenous heritage datasets)
\ **landuse** (landuse datasets)
\ **seause** (seause data incl. Shipping routes, channels etc)
\ **polwaste** (pollution and water management datasets)
\ **admgov** (Government administrative boundaries eg. State borders, LGA)
\ **admother** (other administrative boundaries)
\ **index** (data index – 100k, 250k, 1m datasets)
\ **tenure** (tenure boundaries incl. Reserves, RNEDB, World Heritage)

and under each of these directories:

\ **data** (ARC/INFO native GIS datasets)
\ **doc** (dataset documentation, including AMLs, scripts)

eg. **Q:\ soc_ int \ tenure \ data \ wha_public** (World Heritage Areas: ARC/INFO coverage)

C. SCRATCH or TEMPORARY DATASETS

The intention of the storage system below is for **scratch datasets** or **temporary datsets** (less than 6 months old).

1. Location

Many of these are at present stored on personal "C-drives". This does not, however, allow for regular backup, and restricts access when a staff member may be away for a period, etc.

Ideally, these documents should be stored on a general access drive.

There is less need for these to follow a regular structure, and may often be placed under a directory using the name of the creator.

All datasets, however, should be documented with at least a standard README file (see below).

D. SAMPLE README File

Created by:
Arthur Chapman

Date:
10 Abril 2003

Contents:
GTOPO30 (30-arc-seconds)

Data Citation:
Bliss, N.B., and L.M. Olsen, 1996. Development of a 30-arc-second digital elevation model of South America. In: Pecora Thirteen, Human Interactions with the Environment - Perspectives from Space, Sioux Falls, South Dakota, August 20-22, 1996. http://edcdaac.usgs.gov/gtopo30/README.html

Origin:
LBA Hydronet http://www.lba-hydronet.sr.unh.edu/statgrid/gtopo30/

Custodian:
USGS

Received:
Downloaded from the internet
10 Abril 2003

<if received from an external source, the name and phone number of the contact person should be included>

Files:
g_sa_gtopo30.asc ASCI Grid of Landcover South America above 25 degree S
gtopo30.grd ARC Grid of Landcover for South America above 25 degrees S

Conditions of use:
unknown

Comments:
<if the data was received from a person via email, a copy of the email should be copied and pasted in here>
