COLLECTION DEVELOPMENT POLICY FOR THE NATIONAL GEOSPATIAL DIGITAL ARCHIVE
Version 2.0, November 1, 2006

I. Collection purpose and description of the users

The purposes of the NGDA collection are to archive broad collections of digital geospatial data of the United States, and make it available to users. The primary users of the NGDA are citizens - present and future - of the United States of America. The secondary users of NGDA are all other people who use the Web.

Users of geospatial data are many; the following brief list is a sampling:

- From the university/academic world, undergraduate and graduate students and faculty, especially in those disciplines which deal with geographic areas, e.g., geography, anthropology, history, etc.;
- From elementary and secondary schools, students and teachers looking for information about a specific country or city;

- From the world of business and commerce, commercial vendors of imagery and mapping services; firms that need to know demographic information; realtors;
- Non-profit, non-governmental organizations such as relief organizations; economic and social councils, etc.;
- Persons and firms collecting information about the environment;
- Government agencies - local, state, federal, international.

II. Selection, Evaluation and Prioritization

Geospatial data are produced in large quantities from a wide-ranging group of organizations and entities. Data producers include government agencies, commercial vendors and individuals. It is important to carefully evaluate the needs of the collecting institution in selecting materials to collect and archive. In order to begin the process, the following steps are recommended.

1. Develop awareness of potential collections.

- Contact and/or explore resources at local, regional, state, and federal agencies to find out what data are produced for the area of interest. Attend regional interest group meetings. Sign up for relevant e-lists such as GIS4lib and Maps-L. Read and subscribe to print and online publications focused on GIS data. Consider commercial sources for the area of interest.

2. Check that a potential collection is within the Scope of Coverage (see Section III below).
3. Assign a priority rating to each potential collection using the following questions as a starting point.

- Is collection in scope for the NGDA and for the NGDA collecting institution?
- Is the collection's geographic area of primary importance to the collecting institution?
- Is the collection at risk due to either:
    - The content provider does not archive the content.
    - The file format is becoming obsolete.

4. Obtain resources to collect first priority collections. Resources include collection funds if the data are not free, metadata/cataloging services, server or repository space, and resources to access the data such as computers and relevant software.

5. If resources are available, proceed to second priority collections.

III. Scope of coverage

The scope of collecting is solely in the realm of geospatial digital data. The term, "digital geospatial data," is defined as digital items, displayed as graphics, that are georeferenced or are geographically identified. These are primarily composed of: digital maps; remotely sensed images (e.g., aerial photographs; data collected by satellite sensors); datasets (e.g. shapefiles, layers, geodatabases, etc.); atlases; globes (celestial and terrestrial); aerial views (e.g., panoramas); block diagrams; geologic sections; topographic profiles; etc.

A. Geography
Primary collecting emphasis is on the geographic area of the United States. Data may cover the entire United States including U.S. territories; or data may be focused on a specific state, consortial area (e.g., southern California; metropolitan New York City), county, city, or city neighborhood.

i. United States national or large regional extent
Many subjects listed in Section B are available at a national level. For example, the United States Census Bureau publishes datasets for the geographic regions designated by their decennial census. National datasets available through numerous national government agencies and commercial entities cover many subjects and may be updated on a set schedule, such as census data every 10 years. Some of these datasets may be compiled in such a way to be of a manageable size for viewing, use, and preservation. Examples would include numerous layers in the National Atlas (http://www.nationalatlas.gov/). Others are very large and may require special treatment for archive consideration. For example, the United States Geological Survey produces and continues to update the National Elevation Dataset, served through the National Map (http://www.nationalmap.gov/) website. The full dataset, as of May 2006, is 60 gigabytes of data. It may be the case at this point that such a dataset is too large for any one node to archive and so should be split across a number of archives.

As of Spring 2006, many national agencies producing geospatial data are working out the policies for the archiving of their geospatial resources. Because policies are still being created, it is recommended that important datasets be archived at the local level to provide long-term access to the material.

ii. State, county, or city within the United States.
Datasets at this level are often produced by state, county, or city agencies. Coordination between groups often occurs because of the cost to produce some of the more expensive datasets. For example, this may be the case for aerial photography created at set intervals. Many states have created geospatial clearinghouses for dissemination of popular datasets. Preservation and retention of older datasets is not guaranteed nor often spelled out at the clearinghouse sites.

Basemap data at each level are important to collect including transportation networks, boundary files, water resources, parcel information, and agency-specific data. Statewide clearinghouses are excellent sources for these data. Counties and cities may disseminate their data over the Internet, although it is highly likely one will need to contact these groups directly. Copyright status should be ascertained when collecting data at this level.

iii. ocean-floor coverage: off-shore areas to the limit of the United States' maritime boundary claim.
Data of this type may include multi-beam surveys, sonar readings, electronic nautical charts, and vector data delineating boundary claims. It is created by national and state agencies, as well as academic institutions and research institutes.

B. Subject
Potential types of materials are split across physical and human/cultural geography. The list has been derived from the Library of Congress G Schedule and then modified.

i. Physical geography

- Mathematical geography (surveying and cartography; etc.)
- Physiography (e.g., topography)
- Hydrology
    - Oceanography
    - Rivers and lakes
    - Floods
- Geology, geophysics, mineral resources, and soils
- Climatology
- Biogeography (land use/land cover)
    - Flora
        - General
        - Aquatic
        - Forests and forestry
        - Agriculture
    - Fauna
        - General
        - Aquatic

ii. Human and cultural geography

- Political geography
- Economics
    - Real property; cadastre
- Public lands; ethnic reservations
- Demography; census
- Technology; engineering; public works
- Transportation and communication
- Commerce and trade; finance
- Military and naval geography
- Historical geography

iii. Remotely-sensed images

- Aerial photographs
- Satellite images

C. Date or Chronology

Geospatial data is often subject to versioning. New versions of data layers are released when changes occur in the original dataset. This can be done to correct errors or to account for changes over time. Data may be changed incrementally or on specific dates due to "trigger-events," such as the decennial census.

Once it has been decided which datasets are important to archive, a decision should be made about the frequency with which versioning should occur. For example, Landsat imagery might be collected once a quarter to follow seasonal changes. School district lines may only need to be captured when boundaries are re-drawn. Transportation routes could be updated on a yearly basis. Geospatial data collected immediately after a natural disaster would be collected as soon as it was made available to libraries and/or the public.

D. Format

As part of the NGDA project, only digital materials are collected. The digital data must be accompanied by minimum

core required metadata. Section IV covers metadata recommendations.

Open source, non-proprietary file formats that are either readily manipulated using standard image-processing or geographic-information-system software are preferred (e.g., geotiff, GML). Data in proprietary formats or data whose display is dependent upon proprietary software (e.g., ArcInfo Coverage or GRID) will be dealt with on a case-by-case basis. Some important factors will be how commonly available and used the software is and whether the data may be exported to a non-proprietary format. For example, the ESRI shapefile is a proprietary format, but it is so universally used, the current NGDA nodes will accept data in this format.

Format registries for geospatial data are presently being created to capture relevant representational information about file formats. The Library of Congress has posted the "Sustainability of Digital Formats Planning for Library of Congress Collections" on their Web site (http://www.digitalpreservation.gov/formats/intro/intro.shtml). At present, it contains no information about geospatial file formats, but certainly will in the future. The Global Digital Format Registry (GDFR) is also interested in including geospatial format information (http://hul.harvard.edu/gdfr/). Appendix 2 includes definitions of some of the more common geospatial data formats.

File formats will change over time with new formats being created and older formats falling into disuse. The Archives should continually evaluate if it is possible to migrate older formats into newer ones, decide when and if old formats will be kept, and keep abreast of best practices in the geospatial community regarding file formats.

E. Language
English will be the most frequent language collected due to the nature and focus of the grant; but it is possible that geographic coverage for areas in the U.S. may have text portions in languages other than English.

F. Copyright
Materials without copyright - e.g., public domain data - comprise much of the collection. Acquisition and access to copyrighted data will be governed by an agreement between the NGDA collection node and the data provider. Access to certain data may be restricted for a specified period of time at the request of the data provider. The NGDA nodes do not intend to be dark archives.

G. Exclusions:

- Digital information with a geographic reference, such as a text history of Alabama
- Analog materials
- Straight statistical data not tied to a geographic area
- Data layers external to the United States
- Data that is to remain perpetually "dark"
- The HTML content from Web sites (Geospatial data gathered from Web sites will be collected.)

IV. Metadata recommendations

It is important to collect as much metadata as is feasible. The NGDA recommends a core set of metadata fields with the understanding that it may not be available in all cases. Whether or not content lacking core fields will be ingested into a node is up to the node itself. Significant metadata in rough order of importance:

- Geographic area: This includes information about the extent of the content. Specific node requirements could require coordinates in decimal degrees or words describing the extent ("Arizona counties" qualified by years).
- Type (intellectual content): This includes maps, remote-sensing imagery (aerial photograph; image from satellite), layers.
- Format: This should identify the file types included (e.g., tiff, jpeg, arcexport, shapefile).
- Projection and/or coordinate system.

- Scale and/or resolution. Resolution is often cited when dealing with aerial and satellite imagery.
- Transfer media: This component details the device upon which the data are stored when deposited with the archive (CD-ROM, DVD-ROM, hard-drive, etc.).
- Title: A title is required for each item ingested. A title may also be included for the whole collection.
- Date of information: This would be the date the information was created.
- Issuance information: This includes the issuing agency, the place of issuance, and date of issuance.
- Data Quality information e.g. FGDC metadata elements such as attribute accuracy and completeness report.
- Rights Information e.g. copyright, reproduction of data.
- Date ingested into the archive.
- Contact information for the content provider: This would potentially include a contact person(s), address, telephone/fax, email addresses, or Web site.
- Collection name and description: This may be supplied by the node ingesting the data.
- Controlled-list subject headings: This might be created by the content provider. Hopefully they would provide a full copy, but at the very least a full "bibliographic" citation.
- Other fields: If content provider provides any fields, then they should also include the field name, field definition, and domain (an authority list).

V. Sources for digital geospatial data

Although governmental sources of data are of primary interest, the archive is not to be limited to data generated, or contracted, by federal agencies of the United States, but instead will include digital geospatial data generated by any agency or person. The emphasis will be firstly on nationwide coverage, of any theme, and very often these are generated by federal agencies. It is expected that nodes will collect in this area according to their specific regional and research needs. This may mean that part of the collecting decision is made by the scale of the dataset.

The NGDA nodes will focus on government agencies at all levels and non-profit entities such as professional organizations or environmentally focused non-profits, with a secondary focus on commercial firms, and a tertiary focus on products issued by people.

VI. Coordination and cooperation with other collections

The NGDA's mission is to be a collecting network. As such, collaboration with other institutions is expected and necessary. Because digital geospatial data sets require large amounts of server space, the cooperation of many institutions will be necessary to build an extensive collection. Cooperation agreements written specifically to govern the collecting areas of the partners should include:

- The collecting areas for each participating institution.
- The frequency of updates and versioning for each dataset.
- The length of the agreement.
- The type and level of access to be provided to the collected materials.
- A set interval to review the collection agreement.
- A glossary.

VII. Appendices

Appendix 1: Sample NGDA-node collection development policies
Note: Once this policy has been vetted and finalized, both UC Santa Barbara and Stanford will write Collection Development Policies for their nodes.

Appendix 2: Glossary

Digital Elevation Model
Digital Elevation Models display a 3 dimension-like image of surface elevation by using a raster grid of evenly spaced elevation values. The values are obtained from USGS topographic maps.
Compiled from multiple sources including:
Retrieved June 10, 2006, http://edc.usgs.gov/products/elevation/dem.html
Retrieved June 10, 2006, http://www.landinfo.com/resources_dictionaryAD.htm#d

Digital Orthophoto Quadrangle
A digital orthophoto quadrangle (DOQ) is a computer-generated georeferenced image of an aerial photograph in which image displacement caused by terrain relief and camera tilts has been removed. It combines the image characteristics of a photograph with the geometric qualities of a map.
Retrieved June 9, 2006, from: http://www.usgsquads.com/prod_doqq.htm

Digital Raster Graphic
A digital raster graphic (DRG) is a scanned image of a U.S. Geological Survey (USGS) standard series topographic map, including all map collar information. The image inside the map neatline is georeferenced to the surface of the earth and fit to the Universal Transverse Mercator projection. The horizontal positional accuracy and datum of the DRG matches the accuracy and datum of the source map. The map is scanned at a minimum resolution of 250 dots per inch.
Retrieved June 9, 2006, from: http://topomaps.usgs.gov/drg/

GML
GML is an acronym for Geography Markup Language. An OpenGIS Implementation Specification designed to store and transport geographic information. GML is a profile (encoding) of XML.
Compiled from multiple sources including:
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=gml

Georeferencing
To establish a relationship between page co-ordinates on a planar map and known real-world co-ordinates. Georeferencing allows geographic data sets to be analyzed and compared with one another.
Compiled from multiple sources including:
Retrieved July 31, 2006 from, http://www.geo.ed.ac.uk/agidexe/term?1228
Retrieved July 6, 2006,
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=georeference

Geospatial
Relating to physical features of the earth and their geographic location, including both natural and man-made features. Geospatial data refers to information derived from maps or remote sensing techniques, such as aerial photography or satellite imagery.
Compiled from multiple sources including:
Webster's New Millennium Dictionary of English, Preview Edition (v 0.9.6)
Copyright © 2003-2005 Lexico Publishing Group, LLC
Retrieved June 8, 2006, from: http://dictionary.reference.com/search?q=geospatial&r=66
Directions Magazine: the worldwide source for Geospatial Technology
Retrieved July 26, 2006, from:
http://www.directionsmag.com/press.releases/index.php?duty=Show&id=10412&trv=1

Raster
An image formed using individual dots with color values, called cells (or pixels). Cells are viewed in a rectangular grid with each cell evenly spaced. Aerial photographs and satellite images are examples of raster images used in

mapping.
Raster layers in a GIS system can depict such information as elevation, precipitation, and temperature.
Compiled from multiple sources including:
Retrieved July 28, 2006 from http://data.geocomm.com/helpdesk/glossary-r.html

Remote Sensing
RS is the process of using a recording device not in physical contact with the surface being analyzed to obtain data. Methods include aerial photography and using sensors sensitive to various bands of the electromagnetic spectrum. Equipment can be deployed from aircraft, satellite or space probe.
Compiled from multiple sources including:
Retrieved July 28, 2006 from http://data.geocomm.com/helpdesk/glossary-r.html

Shapefile
Shapefile is the name of the proprietary digital vector storage format created by ESRI Corporation. Shapefiles are used and created in software such as ArcView, Arc/Info, ArcGIS and other widely used GIS software.
A shapefile consists of multiple files that together generate a data layer in a Geographic Information System (GIS). There are three required files that are stored and deployed together in a shapefile:
.shp - the file that stores the feature geometry.
.shx - the file that stores the index of the feature geometry.
.dbf - the dBASE file that stores the attribute information of features.
Other files can be added to the shapefile to carry additional information such as projection and metadata.
Compiled from multiple sources including:
Retrieved June 29, 2006 from, http://en.wikipedia.org/wiki/ESRI_shapefiles
Retrieved June 29, 2006 from, http://walrus.wr.usgs.gov/infobank/programs/html/definition/shapefile.html
Retrieved July 26, 2006, from,
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=shapefile

Vector
Vector data (used in a GIS system) is one method used to store spatial data. Features are defined by their boundaries only and curved lines are represented as a series of connecting arcs. Vector data is expressed as X,Y,Z coordinates. Examples of vector layers include schools (points), street networks (lines), and voting districts (polygons).
Compiled from multiple sources including:
Retrieved June 9, 2006, http://www.geo.ed.ac.uk/agidexe/term?349
Retrieved July 31, 2006, http://data.geocomm.com/helpdesk/glossary-v.html

XML
XML is an acronym for Extensible Markup Language. Developed by the World Wide Web Consortium (W3C), it is a standardized markup language for designing text formats. It enables the interchange of data between computer applications. XML is a set of rules for creating standard information formats using customized tags and sharing both the format and the data across applications.
Compiled from multiple sources including
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=xml

GIS Dictionaries and Glossaries:
http://www.agi.org.uk/bfora/systems/xmlviewer/default.asp?arg=DS_AGI_TRAINART_67/_firsttitle.xsl/87

http://www.fgdc.gov/metadata/csdgm/glossary.html

http://www.gis.com/whatisgis/glossaries.html

http://www.landinfo.com/resources_dictionaryAD.htm

Appendix 3: Collection Levels

0 - Out of Scope
1 - Minimal Level
2 - Basic Level
3 - Study Level
4 - Research Level
5 - Comprehensive Level

Appendix 4: For more information (links and bibliography)
Links

- The Center for International Earth Science Information Network (CIESEN)
  http://www.ciesin.org/

- Digital Curation Centre, United Kingdom
  http://www.dcc.ac.uk/

- Federal Geographic Data Committee (FGDC)
  http://www.fgdc.gov/
    - FGDC Content Standard
      http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html

- National Archives and Records Administration (NARA)
  Requirements for transfer of permanent electronic records, geospatial data:
  http://www.archives.gov/records-mgmt/initiatives/digital-geospatial-data-records.html

- National Digital Information Infrastructure and Preservation Program (NDIIPP)
  http://www.digitalpreservation.gov/

- National Geospatial Digital Archive (NGDA)
  http://www.ngda.org

- National Library of Australia, Preserving Access to Digital Information (PADI)
  http://www.nla.gov.au/padi/index.html

- North Carolina Geospatial Data Archiving Project (NDGDAP)
  http://www.lib.ncsu.edu/ncgdap/

- Maine GeoArchives: A collaborative project between the Maine State Archives and the Geolibrary Board
  http://www.maine.gov/sos/arc/GeoArchives/geoarch.html

- United States National Satellite Land Remote Sensing Data Archive
  http://edc.usgs.gov/archive/nslrsda/

Bibliography
CIESEN links to *Workshop and conference proceedings, articles, and reports*. Retrieved August 1, 2006, from

http://www.ciesin.org/ger/links3.html .

Maryland State Geographic Information Committee. (undated) *Standards for records preservation: Maryland State Government Geographic Information Coordinating Committee*. Retrieved June 20, 2006, from: http://www.msgic.state.md.us/publicat/preserve/ .

McGarva, Guy. (undated) *Curating geospatial data*. Retrieved June 20, 2006:

Filename: CDP_Nov_06.doc
Directory: C:\Documents and Settings\englander\My Documents\files\Project\work groups
Template: C:\Documents and Settings\englander\Application Data\Microsoft\Templates\Normal.dot
Title: COLLECTION DEVELOPMENT POLICY FOR THE NATIONAL GEOSPATIAL DIGITAL ARCHIVE
Subject:
Author: englander
Keywords:
Comments:
Creation Date: 12/12/2006 10:16:00 AM
Change Number: 1
Last Saved On: 12/12/2006 10:30:00 AM
Last Saved By: englander
Total Editing Time: 16 Minutes
Last Printed On: 12/12/2006 10:37:00 AM
As of Last Complete Printing
Number of Pages: 9
Number of Words: 3,457 (approx.)
Number of Characters: 19,711 (approx.)