

May 2007

I. Collection purpose and description of the users

The purposes of the Stanford node of the NGDA are to archive collections of digital geospatial data of the United States of interest to the primary users of Stanford University (students, faculty and staff), and make the metadata and data (the latter may be served out by content providers rather than by Stanford) available to both primary users at Stanford and to primary and secondary NGDA 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, of which there 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. The following steps are normally taken by Branner to build collections in keeping with the needs of its primary users.

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 Branner itself and as a node of the NGDA?
- Is the collection's geographic area of primary importance to Branner?
- 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 firstly of Santa Clara and San Mateo Counties, and of California as a whole, and secondarily of the United States.

i. United States national or large regional extent

General data collected at the national level include demographic (including Census information), base map information (rivers, cities, roads, lakes, etc.), and boundary. Data sources will include the U.S. Census Bureau, the United States Geological Survey, and relevant government Web sites, such as the National Atlas. The National Atlas datasets have often been compiled in such a way to be of a manageable size for viewing, use, and preservation.

ii. State, county, or city within the United States.

GIS digital data is collected at a large scale for the Bay Area, especially Santa Clara and San Mateo Counties, and California with subsequently smaller scale data being collected for the United States. Basemap data at each level is collected including transportation networks, boundary files, water resources, parcel information, and agency-specific data. Data will be collected from local and state agencies, national agencies focusing on the specifically on our collecting region. Coverages created by commercial firms and individuals will also be considered based upon relevancy, copyright, and use restrictions.

iii. Ocean-floor coverage: off-shore areas to the limit of the United States' maritime boundary claim; offshore California constitutes the primary collecting area for Branner. 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.

- For Santa Clara and San Mateo Counties, Branner will collect all subjects listed.
- For California as a whole, Branner will concentrate on the following subjects:
 - topography
 - geology
 - environmental aspects of such areas as oceanography, climatology, economics, etc.

i. Map-format materials (e.g., maps, diagrams, sections, views, profiles)

a. Physical geography

- Mathematical geography (surveying and cartography; etc.)
- Physiography (e.g., topography, bathymetry and hydrography including nautical charts)
- 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

b. 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

ii. Remotely-sensed images

- Aerial photographs
 - primary emphasis: Santa Clara County, San Mateo County
 - other areas and states as appropriate to the needs of Branner's primary users
- Satellite images
 - primary emphasis: California
 - secondary emphasis: other areas and states as appropriate to the needs of Branner's primary users

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 a dataset has been acquired, a decision will be made about the frequency with which versioning will 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. While as a general rule digital data must be accompanied by minimum core required metadata, exceptions will be made for some datasets depending upon their importance to the collection. 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/qdfr/>). The NGDA is building out its own format registry, which will be ingested at a regular basis into the geospatial archives. Appendix 2 of the NGDA CDP defines 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. Branner and the Stanford Digital Repository staff will 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 Branner being located in the United States. It is possible that geographic coverage for areas in the U.S. may have text portions in languages other than English, and especially in Spanish.

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 Branner, as 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. Branner does not intend to be dark archive.

G. Exclusions:

- Digital information with a geographic reference, such as a text history of California;
- Analog materials;
- Straight statistical data not tied to a geographic area;
- Data layers external to California and the rest of 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

The NGDA CDP recommends collecting as much metadata as possible with an understanding that the amount available for any given data set or collection will vary. Branner attempts to capture the metadata fields that the NGDA CDP has identified as core.

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 ("California 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, Branner's collection is not to be limited to data generated, or contracted, by governmental agencies of Santa Clara and San Mateo Counties, the state of California, and the United States, but instead will include digital geospatial data generated by any agency or person. The emphasis will be firstly on Santa Clara and San Mateo County and California coverages, of any theme, and then on the United States or portions of the United States as appropriate to Branner's primary user group.

Branner 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

Within Stanford, Branner coordinates with the Social Sciences Resource Center, which collects census data for the United States.

Branner Earth Sciences Library and Map Collections is a member of the UC/Stanford Map Libraries Group (<http://library.ucsc.edu/maps/ucsmg/>) and works closely with all members of that group. Branner expects to work closely with other NGDA nodes as appropriate.

As noted in the NGDA Collection Development Policy, because digital geospatial data sets require large amounts of server space, the cooperation of many institutions will be necessary to build an extensive collection. Cooperative 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: see NGDA Collection Development Policy