

## Data Formats for Digital Earth

Zitan Chen

*Environmental Systems Research Institute*  
380 New York Street, Redlands, CA 92373, USA  
Tel: (909)793-2853 Fax: (909)793-5953 Email: [zchen@esri.com](mailto:zchen@esri.com)

**ABSTRACT** Digital Earth will be a huge project participating by multi-countries, multi-organizations, and millions of individuals from multi-disciplines. Data from multi-regions at multi-scales to join the project will be from multi-sources, containing multi-languages, representing in multi-data formats and categorizing into multi-security levels.

How to manage this huge volume of spatial data in a practical efficient way for applications developing at global, continental, regional and local scales is a top challenge for current GIS systems and IT.

It is too idealistic to have only one spatial data format to unify all data in Digital Earth at this moment. This article presents three principles for sharable data formats from technical and institutional aspects, in order to reduce potential difficulties for communication insides of Digital Earth.

The first principle is compatible. Sharable data formats should base on similar abstract representation for natural phenomena and processes. They have equal or similar richness level of data contents. Only those formats containing same type and richness of information can be converted each other in two ways communications.

The second principle is extendable. A long life data format must provide convenience to allow more detail and more advanced formats to be developed based on it. It also has to tolerant enough for representing more types of data by only changing or replacing individual module in a format.

The third principle is implementable. None of a format could be existed for long until it is fully supported by practical implementable systems and various application packages.

**KEYWORD** Digital Earth; data format; data share;

### 1. Background

I am the first Chinese to forward Vice President Ai Gole's speech to Chinese GIS society. When my college Dr. Michael Phonix forwarded this speech to me by email, I was very excited and thought China should and will make contribution to this global project. After I load a file of the speech to CPGIS-L, one student asked me that if it is true document, I believed he was too excited, too. I never thought there would be a high level international symposium of Digital Earth at Beijing. During the past year, this project has been seriously discussed, especially in China.

Digital Earth (DE) project is too large comparing with exiting GIS project or other GIS data programs. If we really need a parallel project to reference, it would be interesting to compare DE project in earth science domain to the Human Genome Project (HGP) in biography and medicine disciplines.

- DE is to explore human being's environment; HGP is to explore human being selves.
- Both are only available and proposed during digital information ear, and their results will be stored in computer database systems.
- Both are involving extremely huge data collections and analysis in the history. For example, HGP involves to identify all the estimated 80,000 genes in human DNA and determine 3 billion chemical bases that make up human DNA.

- Both DE and HGP are very fundamental science researches, and will cause many new studies and technology development to stimulate using these significant data bases of human knowledge.

- Both DE and HGP are complex and will be involving with social, legal and institutional issue. International contributions and cooperation play important role.

- HGP project has clean working plan and well-organized international organizations that distribute working tasks to various agencies, including international organizations. All results discovered by each agency will be contributed to the HGP and be published to the worldwide for sharing. HGP was proposed in 1990 for a 15 years plan. New technology made this project develop faster, so it may be completed in 2003.

Digital Earth has not yet achieved the same stage of the Human Genome Project did, due to various reasons. There are still many fuzzy issues regarding to Digital Earth project itself and nobody has answered. One of crucial issues is how to manage and distribute geo-data of Digital Earth.

### 2. Data Management and Distribution of Digital Earth

Based on inherent property of geo-data, Digital Earth will be a huge project participating by multi-countries, multi-organizations, and millions of

individuals from multi-disciplines. Data from multi-regions at multi-scales to join the project will be from multi-sources, containing multi-languages, representing in multi-data formats and categorizing into multi-security levels. Comparing to the HGP project, data collection, management, and integration of Digital Earth is more difficult.

How to manage this huge volume of spatial data in a practical efficient way for applications developing at global, continental, regional and local scales is a top challenge for current GIS systems and IT.

Let us face the facts. There are two types of data: current collected data and data will be collected.

First, many countries already have huge volume of geo-data collected and been managed in their own way. North American countries, Europe countries, Some Asia countries and other many countries have accumulated and collected multi billions dollar value of geo-data. For example, the USGS has owned over twenty two billion dollars of data assets of geo-data. For these existing data, the important task is to provide a program and technology to exchange and to distribute those available data. At least, a mechanize for register all available data for Digital Earth project, including all properties of these data should be in high priority consideration.

Second, there are still much new geo-data that haven't been collected from both developed and developing countries and regions. Some cases are due to that current data are not good enough for modern applications by out off date, or lower quality. For these new data, the crucial issue is to make sure their collection and management in such way that not conflict to be used in Digital Earth project. They should be directly collected and managed in exchangeable and distributable format for more widely used purpose.

It is already noted that, new data collection have moved directly in digital format. New technology, such as remote sensing and GPS, provide powerful tools for the change of traditional surveying, geodesy and cartography. It will be interesting for us to see that some new areas, including the Moon and the Mars, may be available to provide complete enough digital data sooner than some areas on the Earth where large traditional geo-data are still in non-digital format, therefore they can not benefit by advanced digital information evaluation.

operation for polygons became efficient only after polygon topology available. Data format can be roughly seen as an interface between data structure and data representation. There would be very difficult or require extra efforts of developing special algorithms for conversion data between different

### 3. Difficulties to Unify Data Format for Digital Earth

Data formats are fundamental data structure for data collection, management and distribution. For simple systems and small projects, to unify data format is a practical solution. For a region or a country, to unify data format becomes relatively difficult. Furthermore, it might be too idealistic to have only one spatial data format worldwide to unify all data in Digital Earth at present time based on following facts:

- Current existing data are collected and managed in various data formats;
- Digital Earth Data will be collected and managed and updated by multi countries. Each country has own agencies and rules to do their job.
- Mostly, each country only contributes portion of their geo-data to the Digital Earth project, after security restriction and other considerations.

When we face the reality of various data formats from multi data sources, however, we always hope these data can be exchangeable and convertible as wide as applications necessary. Otherwise, various data formats will resist data sharing at least have to spend extra time and efforts for data exchange. In worst case, some of these data in these different formats will be isolated out off Digital Earth project.

### 4. Principles of Data Format Standard

What we should consider to design or to select one data format? Are there any building-in "genes" that make some data formats living longer? This article is trying to explore these properties. It is noted that many spatial data formats have been proposed and announced by various organizations in last two decades, but many of them disappeared and silenced in short time. Considering the cost of heavy labor and huge investment for collecting data in a data format, we definitely hope that the favor data format selected is more popular and can be used longer.

#### 4.1. Data Structure and Contents Level

Data format design or selection must consider its capability of volume and types for spatial information contents to be represented and to be used. Data structure and data contents determine what these data can do and what type of analysis can be performed efficiently. For example, Boolean types of data structure, so as their data format, such as conversion between raster and vector data.

There are two keys: First, one proper data format should cover all necessary contents of spatial information in a system. Second, multi data formats that can be converted each other (two ways

conversions) should have similar level of information contents and data structure.

VPF, SDTS, and ArcInfo format are good examples. They are different data formats designed by different organizations, but their spatial information contents are at same level. Therefore, they can be converted each other in two ways without loss of spatial information. However, DXF and JPEG are different types. They are designed and emphasis on different information contents. They can not be completely converted to each other, regardless how one hardly to try it in programming point of view. Part of useful information will be lost during conversion in one way or another.

Recent GIS development already bring a concept of Metadata into implementation. Metadata contains information of data itself. Many properties, including its resources, parameters, characteristics, and important relating information. It will provide richer information besides data format only for data exchange and distribution.

#### *4.2. Data Format Should be Extendable*

A data format living longer should be easier to adopt new requirement. New geo-features will become necessary for new applications. These new geo-features will introduce new data structures. Then the new data structure needs to be represented by the data format as a new extension, to keep efficient performance of the data format in the new application.

Comparing to original points, lines, polygons and surfaces, many new feature types become more popular. Linear features, networks, solids and temporal changing features, all pop up and need technology to represent them efficiently.

Object-oriented programming introduces two useful concepts for data structure models: encapsulation and inheritance. They have been successfully used in programming and potentially influence to data format. At present time, some comprehensive GIS packages already provided capability of allowing users to establish feature types defined by user themselves. This is powerful extension for data format concept. To allow users to self define new type of data features may be the most important extension among all extensions. This capability will play more important roles, especially in comprehensive new applications.

Any existing selected data format will face pressure to represent these growing data structures and data types. One best way is to allow the selected data format has capability to be expanded for new data structures, and their consequent data format. This criterion should be considered at the

beginning during selection or design of one data format.

#### *4.3. Data Format Must be Executable*

Without strong software tools fully support, none of data format could be popular and livelong. A better-designed data format should provide a friend environment for others to develop software tools to support it. The best data format should attract developers, scientists, vendors and users to build powerful software tools to use the data format. Necessary software tools include databases, application functions, data collection, data manage and data visualization tools. A format without strong popular software to support it can not stand up and be used widely. It will disappear.

It should be emphasized that one necessary part of data collection and management functions is data format conversion. It must be recognized that there are many data formats existed, and already so much information been collected in these multi data formats at different regions. If one selected format that can not convert to and from these existing data formats, consequence is that significant effort to re-collect data in the new data format --- it is huge waste and in some worst cases, it is impossible to recollect data.

Essentially, with or without complete supporting software is one important criterion to evaluate one data format if it will be efficient and practical. Some formats might seem theoretically correct, but it could be a bad selection due to its poor performance and higher costs, since it does not have enough software tools to support it. On other side, if one data format is not successful to attract too much software support after the data format announced several years, one may ask why.

Each new data format must provide correct and precise document to describe its technical details. According to the document, other software then can communicate with this data format. Software vendors need to use the document to develop support software tools. Users have to develop application programs based on the data form at Quite often, many conversion packages must be developed first according to this document to manage all involved data in other data formats.

## **5. Conclusions**

It is a wish to unify all data format into one standard data format at global level. Unfortunately, it is not a reality now. Digital Earth project faces a serious challenge of managing huge volume of geo-data that are from multi resources in different data formats. This article suggests three considerations for selection of spatial data formats for those data

will contribute to DE project: Information contents level; Extension capability; and its implementation capability. All potential data managed by those data formats following these considerations, would be

less trouble to communicate each other, and therefore, will working efficiently and be more popular.