An introduction to GIS: data collection, spatial data models, data manipulation techniques, data storage, and types of data quality

Oral Johnson
University of Technology

***Abstract*—GIS stands for Geographic Information Systems and it is defined from the perspective of tool, technology, service and science. Data collection method includes data capture which constitute primary and secondary data. Map projections are a systematic rendering of points from the curved earth surface onto a flat map surface. The world is seen as discrete objects which represent vector data and as a continuous surface which reflects raster data. Vector uses points, lines, polygons and raster use a cell or grid structure. Geoprocessing is any GIS operation used to manipulate data such as clip or buffer. Automation is the ability to perform repetitive tasks. Geodatabase is a database designed to store, query and manipulate geographic information and spatial data. Data quality is the degree of data excellency that satisfy the given objective, it can be categorized as precision and accuracy. Metadata is information about data, it documents the who, what when, where, how, and why of a data resource.**

Keywords—Map Projection, Data Quality, Geoprocessing, Geodatabase, Metadata, Data collection, Vector data, raster data

Introduction

When Roger Tomlinson coined the term geographic information system (GIS) or the government of Canada he never would have imagined the intensity of debate or the impact the three letter acronym would have (Wright et al., 1997). Chrisman (1999) defined GIS as an organized activity by which people measure and represent geographic phenomena then transform these representations into other forms while interacting with social structures. Shin and Campbell (2011) stated that GIS is used to organize, analyze, visualize, and share all kinds of data and information from different historical periods and at various scales of analysis. In this paper, I critique the various perspectives from which geographic information system is defined. The methods used to capture spatial data will be examined. Eldrandaly (2006) stated that a successful GIS depends in large part of using map projections correctly, this will be then examined. I will also evaluate the conceptual and data models that are the pillars of GIS. Then, I will differentiate between automation and geoprocessing and giving examples of the latter. Lastly I will zoom into aspects of data quality whilst at the same time articulating the importance of metadata and geodatabases in GIS. The motivation to do this research emanated from the class discussions and as a Geomatics major student I felt compelled to do a more profound research. This course will be providing the basics of what to expect and as such this research will deepen my understanding and ensure a solid foundation of the principles and basics of GIS.

What is Geographic Information Systems?

GIS stands for geographic information systems and Arnold (1989) defined it as any manual or computer based set of procedures used to store and manipulate geographically referenced data. Wright et al. (1997) stated that geographic information science is concerned with geographic concept*s*, the primitive elements used to describe, analyze, model, reason about, and make decisions on phenomena distributed on the surface of the earth. Wright et al (1997) also mentioned that science is often used as a generic synonym for research and similarly GIS as a science involve research on a set of basic problems, many of which have been longstanding in traditional disciplines but have been rejuvenated since the development of

GIS. GIS as a tool refers to the use of a particular class of software, hardware, tools such as digitizers and plotters to advance the investigation of a problem (Wright, 1997). Most importantly the tool is separable and independent from the substantive problem and has little to do with the legitimacy of the research.

Two sources of GIS data are hard copy maps and photographs. A map is a physical or conceptual depiction of the characteristics that take place on or near the earth's surface (GIS, 2001). Analogue map is any tangible map production that has a continuous appearance and may be viewed directly (GIS, 2001). Aerial photography is a technique of photographing the earth's surface or features of its atmosphere or hydrosphere (Shin and Campbell, 2011). A camera which act as a sensor is mounted on balloons, aircrafts, rockets, satellites and other spacecraft (Shin & Campbell, 2011). The two broad capture methods are primary and secondary data capture. Both vector and raster models have their primary data capture methods. According to Longley et al. (2005) the dominant form of primary raster data capture is remote sensing which is a technique used to derive information about the physical, chemical and biological properties of objects without direct physical contact. Information is derived from measurements of the amount of electromagnetic radiation reflected, emitted, or scattered from objects (Longley et al., 2005). Sensors that operate through the electromagnetic spectrum are commonly employed to obtain measurements. Aerial photographs as aforementioned are usually collected using analog optical cameras or digital cameras and then later rasterized usually by a scanning film (Longley et al., 2005). Most aerial photographs are collected on an 'ad hoc' basis using cameras mounted in airplanes flying at low altitudes (3000-9000m) (Longley et al., 2005). Primary vector data capture is a major source of geographic data and the two main branches are ground surveying and GPS. Chouan (2018) stated that ground surveying is based on the principle that is 3D location of any point can be determined by measuring angles and distances from other known points. If the coordinate system of the point is known or unknown, all subsequent points can be collected in the coordinate system or from a relative or local one (Chouan, 2018). The Navigation Satellite Timing, And Ranging Global Positioning Systems or NAVSTAR GPS, is a satellite based radio navigation systems that is capable of providing extremely accurate worldwide, 24-hour, 3 dimension (latitude, longitude, and elevation) locational data (Lange & Gilbert, 2005). They're capable of recording position to a high level of accuracy. LiDAR (light dection and ranging) is a relatively recent technology that employs a scanning laser range finder to produce accurate topographic maps (Longley et al., 2005). Geographic data capture from secondary sources is the process of creating raster and vector files and databases from maps, photographs and other hard copy documents (Longley et al, 2005). Scanning is used to capture raster data and scanners are devices that convert hardcopy analog media into digital images by scanning successive lines across a map or document and recording the amount of light reflected from a local data source (Longley et al., 2005). Heads up digitizing, stereo photogrammetry and COGO data entry are the most widely used methods for capturing vector data (Longley et al., 2005). Heads up digitizing and vectorization is the process of converting raster data into vector data; the simplest way to create vectors from raster layers is to digitize vector objects manually straight off a computer screen using a mouse or digitizing cursor (Longley et al., 2005). Photogrammetry is the science and technology of making measurements from pictures, aerial photographs and images (Longley et al., 2005). Measurements are captured from overlapping pairs of photographs using stereo plotters (Longley et al., 2005). It is sometimes the only practical method for capturing geographic data. COGO also refer to as coordinate geometry is a method of capturing data by using survey-style bearings and distances to define each part on of an object (Longley et al., 2010). COGO data are very precise measurements (Longley et al., 2010).

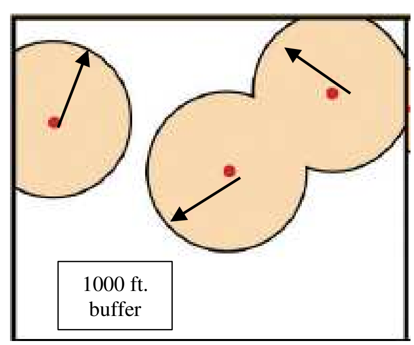
A map projection is a systematic rendering of locations from the curved surface onto a flap map (Eldrandaly, 2006). They're basically mathematical formulas that are used to translate latitude and longitude on the surface of the earth to x and y coordinates on a plane (Shin & Campbell, 2011). Eldrandaly (2006) stated that these map projections used in GIS are based on developable surfaces onto which the latitude and longitude are projected. A developable surface is a surface that can be laid out flat without distortion (Eldrandaly, 2006). The three dominant developable surfaces are cones, cylinders and planes (Eldrandaly, 2006). Points on the earth are projected onto the developable surfaces. This surface is then unrolled to form a flat map (Bolstad 2002 as cited in Eldrandaly, 2006). This process of flattening is not perfect and will cause some distortion and alterations to spatial attributes such as distance, area, shape and direction (Eldrandaly, 2006). Two type of projections are planar and conical projection. With plane projection the spherical plane is projected onto a flat surface. It is also known as an azimuthal projection or a zenithal. This type of projection can be tangent to the globe and it also can be seacant (Kennedy, 2000). The North Pole and South Pole are the most common contact points for GIS databases (Kennedy, 2000). The point of contact can also be at the equator or any point on the earth's surface

Polar

The polar aspects in the diagram above are the simplest form. Parallels of latitude are concentric circles centered on the pole, and meridians are straight lines that intersect at the pole with true angles of orientation (Kennedy, 2000). The planar aspect usually causes longitude lines to converge at the contact point and radiate outwards from the pole like the spokes of a wheel (Kennedy, 2000). Area, shape, and distortions are usually circular around the point of contact. Because of this planar projections accommodate circular regions better than rectangular ones (Kennedy, 2000). In the conic projection latitude and longitude is projected onto a cone either as a tangent or seacant (Kennedy, 2000). A simple projection is tangent to the globe along a line of latitude. This line is called the standard parallel (Kennedy, 2000). The meridians are projected onto the conical surface, meeting at the apex, or point, of the cone. Parallel lines of latitude are projected onto the cones as rings (Kennedy, 2000). The cone is then 'cut' along any meridian to produce the final conic projection, which has straight converging lines for meridians and concentric circular arcs for parallels (Kennedy, 2000). The top of the cone is usually cut to produce a more accurate projection, this is due to distorting increasing away from the standard parallel (Kennedy, 2000).

The two conceptual models of reality that gave rise to the two data models are discrete objects/data where features are seen as objects or entities and continuous spatial variation. According to learn GIS (n.d.) discrete data is data that represents features which exist independently with clear and definable boundaries. The three geometrical primitives: points, lines, or polygons is defined using one or more x,y coordinate pairs called vertices, and are thus described as discrete objects because of their precisely defined locations and boundaries (Conoly & Lake, 2006). Continuous variation or data is the opposite of discrete data, data which does not have clear and definable boundaries but instead makes a blanket of data across a landscape, defined with a scale of values (learn GIS, n.d.). Data such as temperature, precipitation, elevation, slope and aspect are all examples of continuous data and is more suited for raster data. While both data types can have data with either discrete or continuous properties, most often, vector data is described as discrete while raster is described as continuous (learn GIS,n.d)There are two basic data models: vector and raster .In the vector data model, real-world entities are represented using one of three geometrical primitives: points, lines, polygons (Conrolly & Lake, 2006). Each primitive is defined by one or more x, y coordinate pairs called vertices. Points are zero-dimensional vectors defined by two or more coordinate pair, and lines are one-dimensional vectors defined by two or more coordinate pairs (Conrolly & Lake, 2006). Polygons or area are two dimensional objects defined by three or more coordinate pairs. Raster data models us a grid matrix of equally sized cells or pixels to present spatial data. Each cell has a value associated with that represents the attribute status of the object of that location (Conrolly & Lake, 2006). Even location where the variable is not present must be given a value, usually zero (Reynolds, 1992). An advantage of raster data is that it has a relatively simple data structure which also for simple analysis (Davis, 2001). Each cell usually have a single code that is easy to comprehend (Davis, 2001). Because of the simplicity raster data can be manage by any machine, even outdated and slower ones (Davis, 2011). The level of technology and expense can be kept fairly low. In contrast, Davis (2001) stated that one of the major disadvantage of raster is that the files are typically large. Every cell must be coded, even when no data is there, usually zero value (Davis, 2001). Because of this the computer storage needs is high, especially for high resolution formats (Davis, 2001). Each cell tend to generalize in a raster format and because of this the output images are of low resolution and less aesthetically pleasing. On the other hand, because of the accuracy of points, lines and polygons vector data tend to provide a better picture of reality and the output images are more aesthetically pleasing (Shin & Campbell, 2011). The storage capacity is less due to a more compact data structure, file sizes are inherently smaller (Campbell & Chin, 2011). Topology is also inherent which allows for simplified spatial analysis, for e.g. proximity and network analysis (Campbell & Chin, 2011). In contrast, because of the discrete nature of vector data continuous data is poorly stored in vector format (Kumar, 2018). Because of the complex nature of the data structure there are limited shortcuts for storing data like in raster, the location for ever vortex must be stored in the model (Campbell & Chin, 2011). Topology is important but any feature edit requires updates on topology and this is often result in intensive processing, especially when dealing with large datasets (Campbell & Chin, 2011).

Bajjali (2017) defined geoprocessing as the tool and processes used to generate derived data sets from other data using a set of tools. Wang and Selwood (2005) mentioned that a typical geoprocessing takes an input dataset performs an operation on that dataset generating new information return as an output dataset. Automation is the ability to perform repetitive tasks and is one of the shining points of computers (Dixon & Udameri, 2016). Automation includes selecting appropriate geoprocessing tools to operate on a dataset. Automation is useful when a user is conducting similar analysis, running the same sequence of tools for several areas. Two commonly used geoprocessing tools are clip and buffer. Buffering is based on a spatial relationship of proximity. Hehai 1997 as cited in Dong et al (2003) stated that the basic idea of buffering is to create a zonal area of a certain distance around its boundary and to identify the impact and service range to surrounding environment. The area that is within the specified distance is called a buffer zone (Dixon and Udameri, 2016). Shin and Campbell (2011) used the example of a resource manager who wanted to ensure that no areas are disturbed within 1,000 feet of breeding habitat for the federally endangered Delhi sands flower-loving fly situated in Western United States. The habitats for the fly could be denoted as points on Arc map. The buffer tool in arc box will then be selected, inserting the habitat shape file as the input layer and changing the distance unit to feet. Arc map will then generate a buffer zone with a distance of 1000ft around all the habitat points of the species (see figure below).



Peterson (2001) stated that clipping is the GIS operation of extracting a subset of spatial data from a larger data set by selecting only those data located inside (or outside) a selected boundary. Sometimes spatial analysts are working with large geographic areas but wants to work with a more specific area within this larger area. For example a demographer wanting demographic data from the community of Bog Walk in St Catherine. The clip tool could extract bog walk from the parish of St Catherine. Following the clip, all attributes from the preserved portion of the input layer will be included in the output (Shin & Campbell, 2011).

Monomier as cited in Wong and Wu (1996) believed that users of spatial data should be concerned with the quality of data as it can affect the confidence of decision or the success of policies. Devilliers et al. (2010) mentioned that the misinterpretation of spatial data quality issues have already led to a large waste of money a number of suboptimal policies, accidents and deaths. Goodchild (1993) also shared a similar view, he said that errors and uncertainties in data can lead to serious problems, not only in the form of inaccurate results but also in the consequences of decisions made on the basis of poor data. Data quality is essential when a GIS is used to make decisions that, potentially, could adversely impact the data subject (Caprioli et al., 2003). Therefore, data quality should be of high quality as it often translate into successful policies, decisions and projects. Attribute data quality can be assured by using domains in geodatabases, this helps to reduce errors in data entry by eliminating invalid entries (Tennant, 2007). Goodchild 1991 as cited in Kitchin and Tate (2013) defined accuracy as the relationship between a measurement and reality. In other words accuracy refers to how closely a measurement or observation comes to measuring the true value (Humbolt State University, 2019). Precision refers to the degree of detail in the reporting of a measurement (Goodchild 1991 as cited in Kitchin and Tate, 2013). Tiwari et al. (2017) stated that precision also called reproducibility or repeatability is the degree to which repeated measurements under unchanged conditions show the same results.

The desired target or true location is the center or bulls eye. Precision is taking multiple attempts and having them at the same location. This is evident in two diagrams from left as the attempts are in the same positions. Accuracy is having attempts close to the center or true location. The third diagram from left is accurate, even though they're loosely distributed the attempts are close to the center or its true value or location

Baldridge (2012) defined a geodatabase as an alternate way to store GIS in one large file, which can contain multiple point, polygon, and or polyline layers. Tenant (2007) believed that authors in his research focused more on analytical methods in GIS and they rarely discuss the scheme used to organize the data itself. Tenant (2007) created a sample geodatabase structure on Arc GIS online that organized excavation data. Baldridge (2012) said that ESRI is pushing the geodatabase idea, because it is a less messy way of organizing data than having multiple shapefiles. Cioba et al (2011) pointed out that one of the advantages of geodatabase is that it includes domain for attributes. An attribute domain constrains the values that may be entered for a particular attribute which eliminates incorrect or extraneous values. Pittman and Costa (2009) believed that these topological validation rules created by geodatabases reduce errors in data entry and ensure data quality and uniformity. ESRI introduced the file geodatabase with its realease of Arc GIS 9.2. One of its unique advantage is its storage space, one terabyte for each dataset and each file can hold many dataset (Pitman and Costa, 2009). Geodatabase is important as it has great storage potential and able to manage big data and also document relevant metadata making the data easy to catalog, query and identify (Pitman & Costa, 2009).

Wong and Wu (1996) defined metadata as data describing data or data about data (Couloundre, Libourel and Sper, 1998). One type of information that is usually included in metadata is data quality and according to the Spatial Data Transfer standards it should consist of five portions covering lineage, positional accuracy, attribute accuracy, logical consistency and completeness (Wong & Wu, 1996). Metadata is important to any information product document (IPD) as these aspects in spatial data quality is crucial to enable the correct use of spatial data, help potential users decide if particular spatial database is appropriate for specific purpose, and to warn users to what extent they should trust the data (Tenant,2007). Tenant (2007) mentioned that a problem often encountered in the creation and use of archaeological GIS data is the lack of or misidentification of the data source. Attribute fields that records collection methods for field data is often used. Metadata in the form of attribute fields ensure that data within a geodatabase, and its source, are identified. One important element in metadata is data identification and it helps potential users to identify pertinent, critical and important information about any data set. Metadata gives appreciation or description elements of the information in the dataset making it easier for future users to locate the data (Coloundre et al., 1998).

Conclusion

There is no universal definition for the term geographic information systems. Many definitions have been posited by the pioneers. GIS is seen as multidisciplinary as its application and use extends beyond geography. GIS has been defined from various perspective, as a tool, technique, science and service. GIS as a science involve research on a basic set of problems that exists across different disciplines. GIS as a tool is where hardware or software tools is used to advance a problem and most importantly it separable and independent from the problem itself. GIS data can be extracted from different sources and the two most common ones are hard copy maps and aerial photographs. The two broad capture methods are primary and secondary. Primary data is captured from direct measurements whilst secondary data is captured from data that exists already. Raster primary techniques utilize censors/cameras to obtain measurements. For remote sensing the censor operate through the electromagnetic spectrum and aerial photographs which are similar have their censors mounted on planes, balloons and rockets. Primary vector data utilizes satellites through GPS technology, LiDAR technology, and instruments such as total stations through ground surveying to obtain measurements. The main secondary data for raster is a scanner. Heads up digitizing, stereo photogrammetry and coordinate geometry data entry are the most widely used methods for capturing vector data. The earth is a spheroid and is not a developable surface. The transformation of coordinates/points from a curved earth onto a 2D or flat map surface is called map projection. The two conceptual models of reality are as discrete objects and as a continuous surface. Discrete objects such as lakes, ponds have definable boundaries while continuous surfaces have no definable boundaries. The two major data models are vector and raster. Vector data model represents earth features in the form of point, lines, and polygons. Raster data is represented in rows and columns of cells or a grid structure. Vector data usually have better output images than raster data, this is due to the accuracy of the points. Raster data files use up more storage than vector data. Raster has a simpler data structure than vector. Topology is inherent in vector whilst is not in raster data. Geoprocessing operations typically takes a tool to perform an action on an input to create a new dataset. Buffers are used in proximity analysis, it uses distance for spatial analysis. Clip is a tool that extracts a particular area from a larger area and forms a new dataset with the specified area. Data quality is one of the most important aspects of GIS, it has implications on the success of policies and decision making. Precision is the degree to which repeated measurements under unchanged conditions show the same results. Accuracy is how close a measurement is to the true value. Geodatabases are seen as the container in GIS. It is important as it helps to organize and store data. Topological validations used in domains in geodatabses ensure data quality. Metadata is data about data. It is important for future users of data. Description elements are used to locate datasets.

REFERENCES

Accuracy and Precision. (2019). Retrieved from http://gis.humboldt.edu/OLM/Courses/GSP_216_Online/lesson6-2/overview.html

Bajjali, W. (2017). *ArcGIS for environmental and water Issues*. Cham: Springer.

Baldridge, J. (2012). ArcGIS 10: What is a Geodatabase? Retrieved from https://helpwiki.evergreen.edu/wiki/index.php/ArcGIS_10:_What_is_a_Geodatabase?

By, R. A. de, & Huisman, O. (2009). *Principles of geographic information systems: an introductory textbook*. Enschede: The International Institute for Geo-Information Science and Earth Observation (ITC).

Caprioli, M., Strisciuglio, G., Scognamiglio, A., & Tarantino, E. (2003). Rules and Standards for Spatial Data Quality in GIS Environments. In *Proceedings of the 21st International Cartographic Conference (ICC)* (Vol. 21, pp. 1740–1746).

Chouhan, T. S. (2018). *Remote sensing and Gis Gps based resource management*. Jodhpur: Scientific Publishers.

Christman, N. (1999). What does GIS mean? *Transaction in GIS*, *3*(2), 175–186.

Cioban, A., Matei, F., Pop, I., & Rotaru, A. (2011). The importance of geodatabases in a geographical information system. *: Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture*, *68*, 296–300.

Coloundre, S., Libourel, T., & Spery , L. (1998). Metadata And GIS: A Classification of Metadata for GIS. Retrieved from https://www.semanticscholar.org/paper/Metadata-And-GIS:-A-Classification-of-Metadata-for-Coulondre-Libourel/268575a74490f723b7f5d565198a994c1d645d43

Connoly, J., & Lake, M. (2006). *Geographical Information Systems in Archaeology*. Cambridge University Press.

Davis, B. E. (2001). *Gis: a visual approach*. Albany (N.Y.): Onword Press.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS*, *14*(4), 387–400. doi: 10.1111/j.1467-9671.2010.01212.x

Discrete and Continuous Data. (n.d.). Retrieved from https://learngis.org/textbook/section-four-discrete-and-continuous-data

Dixon, B., & Uddameri, V. (2016). *Gis and geocomputation for water resource science and engineering*. Chichester: Wiley.

Dong, P., Rui, X., Cheng, Q., Yang, C., & Zhang, L. (2003). An effective buffer generation method in GIS. In *Geoscience and Remote Sensing Symposium* (Vol. 6). doi: DOI: 10.1109/IGARSS.2003.1295244

Eldrandaly, K. A. (2006). A COM-based expert system for selecting the suitable map projection in ArcGIS. *Expert Systems with Applications*, *31*(1), 94–100. doi: 10.1016/j.eswa.2005.09.008

GIS in fisheries management . (2001). Retrieved from http://webco.faocopemed.org/old_copemed/reports/gis/maltaCourse/maps.pdf

Goodchild, M. (1993). Data models and data quality: Problems and prospects. Retrieved from http://www.geog.ucsb.edu/~good/papers/192.pdf

Kennedy, M. (2000). *Understanding Map projections*. Environmental Systems Research Institute.

Kitchin, R., & Tate, N. J. (2013). *Conducting research in human geography: theory, methodology and practice*. London: ROUTLEDGE.

Kumar, G. (2018). *G I S principles and practices* . Educreation Publishing.

Lange, A., & Gilbert, C. (2005). Using GPS for GIS data capture . In *Geographical Information Systems: Principles, Techniques, Management and Applications, 2nd Edition, Abridged* (pp. 467–476). John Wiley & Sons.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic Information Systems And Science*. John Wiley & Sons Ltd.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2010). *Geographic information systems and science* (3rd ed.). Hoboken: John Wiley & Sons.

Penant, E. (2007). A Sample Geodatabase Structure for Managing Archaeological Data and Resources with

ArcGIS. *Technical Briefs In Historical Archaeology*, *2*, 12–23.

Peterson, D. M. (2001). *Development of Archean lode-gold and massive sulfide deposit exploration models using geographic information system applications: targeting mineral exploration in northeastern Minnesota from analysis of analog Canadian mining camps*. Ann Arbor, MI: University of Minnesota.

Pittman, S. J., & Costa, B. (2010). Linking Cetaceans to Their Environment: Spatial Data Acquisition, Digital Processing and Predictive Modeling for Marine Spatial Planning in the Northwest Atlantic. *Spatial Complexity, Informatics, and Wildlife Conservation*, 387–408. doi: 10.1007/978-4-431-87771-4_21

Shin, M. E., Campbell, J. (2011). *Essentials of geographic information systems*. Boston, MA: Flat World.

Tang, W., & Selwood, J. (2005). *Spatial portals: gateways to geographic information*. Redlands: ESRI Press.

Tiwari, V., Tiwari, B., Thakur, R. S., & Gupta, S. (2017). *Pattern and data analysis in healthcare settings*. Hershey, PA: Medical Information Science Reference, an imprint of IGI Global.

Wong, D., & Wu, C. (1996). Spatial metadata and GIS for decision support. *Proceedings of HICSS-29: 29th Hawaii International Conference on System Sciences*, 557–566. doi: 10.1109/hicss.1996.493251

Wright, D., Goodchild, M., & Proctor , J. (1997). Demystifying the Persistent Ambiguity of GIS as "Tool" Versus "Science". *The Annals of the Association of American Geographers*, *87*(2), 346–362.

Wyatt, P., & Ralphs, M. (2003). *Gis in Land and Property Management*. London: Taylor & Francis.