Monitoring of Technology Adoption Using Web Content Mining of Location Information and Geographic Information Systems: A Case Study of Digital Breast Tomosynthesis

abstract

Purpose To our knowledge, integration of Web content mining of publicly available addresses with a geographic information system (GIS) has not been applied to the timely monitoring of medical technology adoption. Here, we explore the diffusion of a new breast imaging technology, digital breast tomosynthesis (DBT).

Methods We used natural language processing and machine learning to extract DBT facility location information using a set of potential sites for the New England region of the United States via a Google search application program interface. We assessed the accuracy of the algorithm using a validated set of publicly available addresses of locations that provide DBT from the DBT technology vendor, Hologic. We quantified precision, recall, and F1 score, aiming for an F1 score of $\geq 95\%$ as the desirable performance. By reverse geocoding on the basis of the results of the Google Maps application program interface, we derived a spatial data set for use in an ArcGIS environment. Within the GIS, a host of spatiotemporal analyses and geovisualization techniques are possible.

Results We developed a semiautomated system that integrated DBT location information into a GIS that was feasible and of reasonable quality. Initial accuracy of the algorithm was poor using only a search term list for information retrieval (precision, 35%; recall, 44%; F1 score, 39%), but performance dramatically improved by leveraging natural language processing and simple machine learning techniques to isolate single, valid instances of DBT location information (precision, 92%; recall, 96%; F1 score, 94%). Reverse geocoding yielded reliable geographic coordinates for easy implementation into a GIS for mapping and planned monitoring.

Conclusion Our novel approach can be applicable to technologies beyond DBT, which may inform equitable access over time and space.

Clin Cancer Inform. © 2018 by American Society of Clinical Oncology

INTRODUCTION

Adoption of new medical technologies is part of the well-known stages of diffusion, typified by innovators, early adopters, early majority, late majority, and laggards.^{1,2} Adoption of a particular technology occurs over geographic and temporal boundaries, with variation in uptake into health care markets and the populations they serve. Monitoring adoption can be an invaluable resource to health care providers and researchers as it can help ascertain which geographic regions and/or populations do not have access to new, beneficial medical technologies. Over

time, monitoring of technology diffusion can reveal potential areas of oversupply. Thus, monitoring technology adoption in real, or near-real time, allows us to analyze availability both serially and in snapshots of time, which may help drive efficient and equitable access to the important health care advances.

Monitoring the adoption of health care services is typically limited one or more key ways—retrospective methods, limited geographic extents, issues of ascertainment, or reliance on structured data.³ Data sources from which to identify health services, particularly billing claims with

Tracy Onega
Dharmanshu Kamra
Jennifer Alford-Teaster
Saeed Hassanpour

Author affiliations and support information (if applicable) appear at the end of this article.

The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the National Institutes of Health.

Corresponding author: Jennifer Alford-Teaster, MA, MPH, Geisel School of Medicine, Dartmouth College, 1 Medical Center Dr, Lebanon, NH 03756; e-mail: jennifer.a.alford-teaster@dartmouth.edu.

service-specific codes, are often only available to researchers from one to several years after claims were filed, thus limiting timely monitoring.³ In the United States, the only data with a fully national extent from which to capture medical technologies and health services are Medicare claims.⁴ These data are usually available with a 2-year lag and include approximately 98% of individuals age ≥ 65 years; thus, younger adults and children are not represented. Timelier and population-representative data can be accessed, such as from individual health care organizations or private insurance companies, but these are limited in geographic extent. In addition, because the ascertainment of health services is typically based on billing codes, new technologies can only be identified once an approved code is established, which occurs after the technology is already in use, thus limiting timely data. Full ascertainment of a new technology in a timely manner may be made *via* medical chart surveillance, but currently this is only feasible on a limited scale, such as in a single hospital. To date, collection of data to monitor the adoption of new medical technologies relies on structured data, as in claims files or electronic health records. There is no canonical metadata available publicly from health care entities that document services or service sites, which, if in place, would create the gold standard for adoption and diffusion monitoring. With the expansion of informatics tools that allow unstructured data, such as free text, to be mined for discrete data elements, the limitations in our ability to monitor technology adoption can largely be overcome.

Although currently there are no spatiotemporal methods with which to explicitly monitor technology adoption that are both timely and of broad geographic extent, informatics and geospatial methods can be harnessed for this purpose. Specifically, Web content mining for publicly volunteered location information⁵⁻⁸ of new technologies can be combined with geographic information systems (GIS) to identify medical technologies in near-real time without regard to geographic extent.

Using digital breast tomosynthesis (DBT) as the example technology, we created a tool that identified this new technology across the New England region of the United States. DBT, also called three-dimensional mammography, is a radiographic method of mammography that

produces three-dimensional images of the breast by using low-dose X-rays at several different angles.⁹ It is rapidly replacing traditional two-dimensional mammography, although the extent of adoption is not known nationally. With DBT having received its own billing codes in 2015 and approval for Medicare reimbursement¹⁰ and other payers, the radiology community expects a dramatic uptake of DBT into all breast imaging facilities in the United States.¹¹

Our method makes use of the Google search application programming interface (API) to identify the presence of the technology at a particular facility through Web content mining and Google Maps API to resolve its location into two-dimensional coordinates. Web content mining uses natural language processing and information retrieval techniques, including data mining and machine learning. Web content mining is widely applied in real time by Web search engines to find and summarize heterogeneous content. Our methods for the automated discovery and extraction of instances of DBT and associated locational—that is, address—information within Web pages, including hyperlinks, were based on prior similar work^{12,13}; however, these methods have only been applied to a limited extent within health and health care and have not been used to spatially ascertain imaging modalities.

The objective of the current work was to describe an approach that combines natural language processing, machine learning, and GIS to locate new breast imaging technology in near-real time. We describe the application of Web-mined publicly available geolocation data within a GIS environment, which has the potential to address multiple issues within health resources allocation and equity in access to new technologies. Although we use the case of DBT as an exemplar technology, the basic approach we have developed for this tool is expected to be applicable to other health care technologies to aid health care systems, policymakers, and researchers.

METHODS

Study Area and Data Sources

Our study area included the New England region states within the United States—Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island. Although our method is largely not scale dependent, we chose a

limited region for the development of the tool because of the need for validation with existing data to assess algorithm performance, as will be described. We required data sources for two major purposes: to create a whitelist of domain names to configure the search engine, and to validate extracted information and assess its accuracy. The whitelist was created using the US Food and Drug Administration (FDA) list of all certified mammography facilities in the United States,⁹ which contained facility names, street addresses, cities, states, and postal codes. FDA data do not indicate the presence of DBT, rather these data only cover two-dimensional mammography facilities; thus, DBT locations cannot be identified through these data. Of note, DBT can only diffuse or be adopted by existing certified two-dimensional mammography facilities. Validation data came from the DBT vendor, Hologic (Marlborough, MA), which was the only approved vendor in the United States at the time of our study, and hosts a Web site from which location information is available for its DBT technology.¹² Facility names, street addresses, cities, states, postal codes, and telephone numbers are provided for a user-specified geographic extent. Using this feature, we manually derived a list of all DBT facilities for the New England study region, as of March 2014, with an update in April 2015, to be used to assess algorithm accuracy.

Geoinformatics Data Acquisition Architecture for Publicly Available Location-Based Web Content Mining

The architecture of our system with which to perform the Web content mining had four main components: the whitelist, Google search functionality, database/location extractor, and d) a Python script that was designed for the integration of the aforementioned components. The whitelist formed the entry point into our system and consisted of a list of radiology and hospital facility names and domains. The list was generated from FDA-certified mammography facilities as well as by programmatically scraping public-domain Web sites of US hospitals and state-licensed facilities.¹⁴⁻¹⁶ The whitelist was grouped by state such that an independent search engine was made for each state. We used a different search engine for each state because the Google API provides a limited output for the free version of the API. The Google search engine

```
params : | tomosynthesis OR "3d mammography" OR "3-d  
mammography"-filetype:pdf-filetype:csv |
```

This query requests the search engine to search the whitelist domains for terms "tomosynthesis," "3d mammography," or "3-d mammography", and exclude PDF or CSV documents. The result is a paginated list of results, similar to what we see in Google search results.

Fig 1. Sample search query using key terms in the Google search engine.

was used to query each Web site in the whitelist for key terms to identify DBT. Key terms were taken from known DBT Web sites, practicing radiologists, and the literature. A sample search with key terms is shown in [Figure 1](#). Series of queries were run to exhaust the key term combinations and permutations for every domain in each whitelist to identify domains that contained DBT. Using the Google Maps API,¹⁵ we resolved the location of the DBT facility and stored the associated location information in a data file with the following fields: facility name, address, city, state, postal code, phone, latitude, longitude, Web link, domain, verification of Web site, and any notes. Despite the availability of Google Maps for manual searches of location information, our geoinformatics application provides a unique and distinct set of advantages for DBT location discovery: completeness of ascertainment when using a single line of search terms in a manual search versus full compilation of validated search terms—for example, "DBT," "tomosynthesis," "digital breast tomosynthesis," "3-D mammography," etc; geographic scale of coverage (infeasible to do a manual search nationally, and for multiple states; laborious and unautomated); reproducible iterations over time (and space as in the preceding point; and ability to integrate automatically in a GIS to map and combine with other data, such as Census, to understand the underlying population context of the technology location information. Thus, the advantages of the method are related to completeness of ascertainment, reproducibility, ancillary data integration and automaticity.

Natural Language Processing and Machine Learning Approach

Natural language processing was used to identify pages with facility address and address information that was extracted via text qualifiers. For example, text strings that included integer numbers, street names, and suffixes denoting

address—for example, “St,” “Street,” “Rd,” “Road,” “Ln,” “Lane,” “Dr,” “Drive,” etc—and five-digit ZIP codes, available via lookup table, were identified as addresses. Validation of address extraction was possible using both the Google API and ArcGIS. A machine learning algorithm was applied on the basis of a simple decision tree approach. Specifically, when no whitelist of potential facilities was used, we created a script that searched for both our DBT keywords, as well as address identifiers, and applied a series of decision rules—for example, contains DBT keywords plus location identifiers; if yes: contains health care facility identifier; if yes: does NOT contain both parent and child pages/URLs, etc. This generated a list of ZIP codes with corresponding false and true positives. We then applied a script that created annotations to serve as input into the Google search API, which created a de facto whitelist. Disambiguation of Web pages, entries, and location information was performed as in the initial Web content mining approach in which we had a whitelist a priori.

The architecture of our Web content mining system relied on a Python script to integrate and execute the above components. The script contained the following functions: extract (extracted domains from the whitelist and groups by state), configure (created a Google search engine configuration for each state’s whitelist), annotator (created annotations to serve as input to the Google search API), and location resolver (generated a CSV file with facility location information and fed to the Google Maps API for reverse geocoding to generate latitude and longitude). The CSV file with latitude, longitude, and facility attributes was then exported to the ArcGIS geodatabase¹⁷ environment where other spatial layers, such as road networks, geopolitical boundaries, and census data, were incorporated for geoprocessing, spatiotemporal analyses, and geovisualization of DBT diffusion using the ArcGIS Desktop application. The geoinformatics data acquisition system is schematically shown in [Figures 2A and 2B](#).

Accuracy Assessment

To evaluate the accuracy of our Web content mining system, we reviewed a random sample of 100 of the top 5,000 hits in our initial search and extracted the output manually to measure true positives (TPs) and false positives (FPs). In this work, we used a combined set of FDA and

Hologic data as the gold standard. Facilities in the Hologic data are true positives, those in FDA data (two-dimensional mammography only), but not in Hologic are true negatives. On the basis of our derived DBT data, DBT facilities will be test positives, and facilities not in our database will be test negatives. We calculated precision (1), recall (2), and F1 score (3), which are standard metrics for information retrieval evaluation¹⁷:

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

$$F1\text{-score} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \quad (3)$$

The development of the presented system was iterative on the basis of our validation results. An F1 score of $\geq 95\%$ was decided as an acceptable performance for this system from the start of this project.¹⁸ We leveraged different described Web and text-mining techniques to tackle the computational and informatics challenges of the current work and to achieve an acceptable performance ([Table 1](#))

We also performed a subanalysis in which we did not use a whitelist and assumed no a priori knowledge of possible DBT locations. Instead, we used our Python Web scraper to crawl through the Web identifying keywords that denoted hospitals, imaging facilities, and the DBT key terms. In this exploratory approach, we additionally included ZIP codes to the Web crawler search to help differentiate location-based instances of DBT mention from purely informational. We report only on the whitelist-based approach in this work, as shown in [Figure 2A](#).

RESULTS

We were able to develop a geoinformatics platform for data acquisition that mined publicly available location information from the Web for DBT facilities that could be imported into a geodatabase for analysis and visualization within a GIS environment. The Web content mining algorithm required several iterations to yield reasonable performance. The performance was assessed by precision, which is the number of correct results divided by the number of all returned results; recall, which is the percentage

Fig 2. System architecture for a geoinformatics data acquisition approach using Web content mining (A) to ascertain new imaging technology locations in near-real time for integration into a geodatabase for geospatial analysis and visualization (B). API, application programming interface; DBT, digital breast tomosynthesis; GIS, geographic information system; mammo, mammography.

of all relevant documents returned by the search; and F1 score, which is the weighted average of precision and recall.¹⁷ Our first pass yielded low precision, recall, and F1 score (35%, 44%, and

39%, respectively; Fig 3A). We improved the algorithm with the proposed machine learning techniques (Table 1), ultimately achieving precision of 92%, recall of 96%, and F1 score of

Table 1. Summary of Key Challenges and Strategies With Which to Address Them in the Geoinformatics Data Acquisition Platform

Informatics Challenge to Geolocation Web Content Mining	Strategy to Address Challenge
Results from nonfacility Web sites; these mention DBT, but may be scientific articles, news articles, blogs, etc.	Created a whitelist of domains to be queried by the search API.
Multiple results yielded from the same Web site; search results are returned using a page-rank algorithm, but multiple pages within a single Web site may be ranked independently from one another.	Applied our own ranking on the basis of the likelihood of the facility providing DBT; each link was grouped in a Python dictionary with domain as key value and associated URLs inside receiving lower ranks.
The same facility results in multiple pages.	Pot hoc parsimony.
Search engine limits via third parties (Google, Bing); a limit of 1,000 results will be returned from a given keyword when using their free API.	Split our search engine into multiple, smaller pieces—in our case, by state.
Extracting location of the facility from Web sites and only capturing one unique address per facility.	Reverse geocoding using Google Maps API; facilities with multiple locations listed on their Web site may have some error.

Abbreviations: API, application programming interface; DBT, digital breast tomosynthesis.

94% (Fig 3B). Although the F1 score of 94% fell below an accepted threshold of 95%, the algorithm performed well nonetheless, as it was close to this threshold.

After extraction of Web pages that mentioned DBT, we resolved the result to a location. To extract the address of the facility and determine its latitude and longitude, we reverse geocoded using the Google Maps API. We found Google Maps to spatially correspond closely, although not exactly, with geocoding in ArcGIS using their base layers. This is a rounding issue, possibly a result of differences in the underlying street data sets or the data output format used for

the geocoding processes.¹⁹ Figure 4 shows the results of the Web-based, automatically geocoded DBT locations within a GIS environment. The geoinformatics platform for data acquisition was able to discern from the collection of US mammography facilities that had adopted DBT, having been validated by an a priori set of validated DBT locations.

DISCUSSION

In the current work, we have developed a novel geoinformatics platform for data acquisition to address a health services research issue related to the timely monitoring of new medical

Fig 3. (A) Accuracy assessment for initial Web content mining algorithm. (B) Accuracy assessment for Web content mining algorithm after applying machine learning techniques. DBT, digital breast tomosynthesis.

A			
Web content mined results			
Gold standard validation category	Test Positive (DBT site retrieved)	Test Negative (DBT site not retrieved)	
Positive (Hologic DBT site)	True positive 22	False negative 28	50
Negative (not Hologic DBT site)	False positive 40		
	62		
Precision: 35% Recall: 44% F1 score: 39%			
B			
Web content mined results			
Gold standard validation category	Test Positive (DBT site retrieved)	Test Negative (DBT site not retrieved)	
Positive (Hologic DBT site)	True positive 48	False negative 2	50
Negative (not Hologic DBT site)	False positive 4		
	52		
Precision: 92% Recall: 96% F1 score: 94%			

Fig 4. Regional digital breast tomosynthesis locations derived from Web content mining in conjunction with training and validation data layers. Reverse-geocoded locations were then imported into a geodatabase.

technology adoption. The platform was based on Web content mining using natural language processing and machine learning to derive geolocations of publicly available addresses of DBT, which were then spatially referenced in a GIS environment. The performance of our final system was only marginally inferior to that of a widely accepted threshold for a high degree of accuracy. Automated extraction and geocoding, with subsequent importing to a geodatabase for use in the ArcGIS system, resulted in an efficient and accurate way to map, analyze, and report on the basis of the spatial distribution of the

Web-based location, as well as producing relevant characteristics of the underlying population. Machine learning techniques and computational capacity are important elements of mining location information from the Web without having an a priori set of potential domains.

The novel aspect of this study was in combining two common methodologic approaches, natural language processing and machine learning and GIS, to address health services monitoring. Ascertaining publicly available geolocations of DBT was most efficient when working from a

prespecified list of domains from facilities that could potentially have the technology. We found few false negatives, and these only occurred in two conditions: a hospital or imaging facility that provided the service but did not advertise it on the Web, or the facility Web site was inadvertently omitted from our whitelist. FPs were notably reduced when we integrated natural language processing with machine learning techniques within the system; however, we note that the most appropriate application of this method is in the context of new services/devices that are subject to market forces such that health care facilities and/or systems act in a competitive environment, as well as to inform their patient populations. Development of this geoinformatics platform in the context of DBT was facilitated by several factors that are specific to this technology—validation sources from two-dimensional FDA mammography facilities, uptake that was likely limited to that denominator of facilities that already provide two-dimensional mammography, and the vendor's (Hologic) location search tool for limited areas, with which we could also validate DBT extraction accuracy. In addition, this method of timely geolocation of new technology measures potential access, and does not account for other barriers, such as cost, transportation, health beliefs, etc.

Several challenges in developing the system created limitations. First, the choice of Web crawler affects both cost and scalability. The publicly available Google search API has a limit of 1,000 records returned per search engine instance; thus, this limitation required us to create multiple search engine instances. This limits scalability when a large number of identified records is expected or possible. Two ways around this limitation are use the commercial version of Google search API, which has some cost, but is not limited in yield; and use another publicly

available Web crawler. For example, Nutch¹⁸ is an open-source Web crawler that can be used to replace Google search. Theoretically, for the objective of medical technology diffusion surveillance, one can leverage Nutch to create a search engine that runs on a local server cluster which could index all the health care facilities, thereby negating the need of a whitelist or state-by-state search engines; however, to find the facility domains, which are scattered randomly across the internet, it would take massive computational capacity.

One of the major drawbacks of this work is its need to disambiguate the location of the hospital. The need for disambiguation may arise when there are facilities with multiple locations, but DBT is provided at only one of them. This challenge may have contributed to our slightly below desirable F1 score (94%). Additional refinement of machine learning algorithms is needed to improve correct locational attribution of DBT when a facility has multiple locations.

In conclusion, Web-based publicly available location information related to medical technology adoption can be harnessed to provide more spatially scalable and timely monitoring of which areas and populations have geographic access to new technologies. Short of authoritative, publicly reported services and provision locations and a geoinformatics approach that integrates natural language processing and machine learning into a GIS environment can be widely applicable to health care technologies beyond the current example of DBT. Understanding spatiotemporal patterns of new medical technologies provides near-real time information that could be used to promote equitable access and efficient distribution.

DOI: <https://doi.org/10.1200/CCI.17.00150>

Published online on ascopubs.org/journal/cci on June 14, 2018.

AUTHOR CONTRIBUTIONS

Conception and design: All authors

Financial support: Tracy Onega

Administrative support: Tracy Onega

Provision of study materials or patients: Tracy Onega

Collection and assembly of data: Tracy Onega, Dharmanshu Kamra, Jennifer Alford-Teaster, Saeed Hassanpour

Data analysis and interpretation: Tracy Onega, Jennifer Alford-Teaster, Saeed Hassanpour

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about

ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Tracy Onega

No relationship to disclose

Dharmanshu Kamra

No relationship to disclose

Jennifer Alford-Teaster

No relationship to disclose

Saeed Hassanpour

No relationship to disclose

ACKNOWLEDGMENT

We thank Amar Das, Steven Andrews, Craig Ganoe, and Xun Shi for support of this work. We also thank the reviewers of this manuscript for providing thoughtful and valuable input, which has strengthened our reporting of this work.

Affiliations

Tracy Onega, Dharmanshu Kamra, Jennifer Alford-Teaster, Saeed Hassanpour, Geisel School of Medicine, Dartmouth College; and **Saeed Hassanpour**, Dartmouth College, Hanover, NH.

Support

Supported by The Dartmouth Clinical and Translational Science Institute, Grant No. UL1TR001086 from the National Center for Advancing Translational Sciences, National Institutes of Health.

REFERENCES

1. Brown L, Cox K: Empirical regularities in the diffusion of innovation. *Ann Assoc Am Geogr* 61:551-559, 1971
2. Rogers E. *Diffusion of Innovations*. New York, NY, The Free Press, 1995
3. Onega T, Alford-Teaster J, Andrews S, et al: Why health services research needs geoinformatics: Rationale and case example. *J Health Med Inform* 5:176, 2014
4. Centers for Medicaid and Medicare Services: CMS data navigator. <https://dnav.cms.gov/>
5. Sui D, Elwood S, Goodchild M (eds): Volunteered geographic information, the exaflood, and the growing digital divide, in: *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Dordrecht, the Netherlands, Springer, 2013, pp 1-12
6. Brownstein JS, Freifeld CC, Madoff LC: Digital disease detection: Harnessing the Web for public health surveillance. *N Engl J Med* 360:2153-2155, 2157, 2009
7. Mooney P, Corcoran P, Ciepluch B: The potential for using volunteered geographic information in pervasive health computing applications. *J Ambient Intell Humaniz Comput* 4:731-745, 2013
8. Eysenbach G: Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res* 11:e11, 2009
9. US Food and Drug Administration: Mammography facilities. <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMQSA/mqsa.cfm>
10. American College of Radiology: <https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Breast-Tomosynthesis>
11. Aunt Minnie: Breast tomosynthesis in the U.S.: Two years later. <http://www.auntminnie.com/index.aspx?sec=ser&sub=def&pag=dis&ItemID=103907>
12. Hassanpour S, O'Connor MJ, Das AK: A semantic-based method for extracting concept definitions from scientific publications: Evaluation in the autism phenotype domain. *J Biomed Semantics* 4:14, 2013
13. Hassanpour S, O'Connor MJ, Das AK: A framework for the automatic extraction of rules from online text. Presented at RuleM, Barcelona, Spain, July 19-21, 2011
14. Wikipedia: Lists of hospitals in the United States. https://en.wikipedia.org/wiki/Lists_of_hospitals_in_the_United_States
15. US Hospital Info: United States hospitals and medical centers. <http://ushospital.info>

16. Department of Health and Human Services: Health facilities licensed under RSA 151. <http://www.dhhs.nh.gov/oos/bhfa/documents/licensedfacilities.pdf>
17. Google: Google Maps API: Web services. <https://developers.google.com/maps/Web-services/>
18. Powers DMW: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Int J Machine Learning Technology. 2:37-63, 2011
19. Esri: What is a geodatabase? <http://desktop.arcgis.com/en/arcmap/10.3/manage-data/geodatabases/what-is-a-geodatabase.htm>