Undefined 1 (2014) 1–5
IOS Press

Extracting Human-Level Knowledge from Big Numerical Spatial Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Matthew Klawonn [a,*,**], Paulo Pinheiro [a] and Deborah McGuinness [a]
[a] *Tetherless World Constellation, Department of Computer Science, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, USA*
E-mail: {klawom, pinhep,}@rpi.edu, dlm@cs.rpi.edu

Abstract. "Big data" has been thrust into the public consciousness as of late, carrying with it sometimes extravagant promises of innovation and benefits for everyone. Certainly there appears to be no shortage of data in sight, yet following through with the promise and potential of big data poses a number of challenges. One significant roadblock to making use of big data is the time and expertise required to perform analysis and derive products from the data. The more the process of extracting spatial knowledge based on human-level concepts, e.g., lakes, valleys, and relations, e.g., near, after, can be automated through the use of computational data analysis, the more time there is for humans to ask important questions and draw conclusions from the data. Here a methodology for extracting human-level knowledge from spatial data is presented alongside an automated means of encoding the derived products into Linked Data for use by a question answering system. The methodology is extensible since it provides a spatial framework that can be used to aggregate any new additional knowledge that may become available as new techniques for extracting knowledge from spatial data is made available.

Keywords: Linked Data, Geonames, The Jefferson Project, GeoSPARQL, Invasive Species

1. Introduction

Geoinformatics is just one of many fields that can benefit from an abundance of data. Traditional geographic information systems (GISs) are well equipped to handle data based on coordinates on some map projection. However, GISs are not particularly well suited to be useful in question answering systems. One challenge is that the data formats are not always easily usable with the wealth of knowledge available in linked data formats. Further, it can be difficult to combine raw geospatial data with the knowledge they conceal. For example, consider a project that takes surface temperature measurements at various points in the Atlantic Ocean. Using current methods, the data and the derived analysis would likely be stored in different places, in potentially different formats, and certainly not in a format friendly to QA systems, i.e linked data or natural language texts.

In order to effectively use geospatial data with question answering systems, there must be a means of performing analysis and automatically transforming data and derived products into formats compatible with QA systems. We have developed a process by which we can take bathymetry data from its raw form and convert it to linked data that contains "human level geospatial terms" that can be useful in geospatial question answering tasks. Our methodology provides a logical connection between spatial data and human-level knowledge.

*Corresponding author. E-mail: editorial@iospress.nl.
**Do not use capitals for the author's surname.

Specifically we make the following contributions. First, we use clustering algorithms to split spatial data into regions. We automatically combine these regions with well known named geographic entities to create a context for the regions which is easily digestible by humans. Secondly, we have developed a means of performing ecological niche modeling using geospatial data, and placing this analysis in the context of regions and named geographic entities. Section 1 describes our knowledge extraction and formatting pipeline. Section 2 explains the motivations for this particular case study, and how the approach taken will help solve the problems presented by the case study. Section 3 describes the data analysis techniques used, section 4 overviews the methods used to transform the derived products into geospatial knowledge encoded as linked data, which are exploited through GeoSPARQL, and the last sections analyze the approach's limitations as well as its future.

2. Invasive Species Monitoring and Prediction: A Case Study

2.1. Jefferson Project

The Jefferson Project is "a three-year, multi-million dollar collaboration with the goal of understanding and managing complex factors -including road salt, storm water runoff and invasive species- threatening [Lake George]."[1] As a part of the Jefferson Project, LiDAR and Sonar data were obtained from Lake George in order to accurately model the lakebed and surrounding area. Developed in the early 1960s, the use of Lidar in geographical and geological applications has been increasing since its invention. "LIDAR (Light Detection and Ranging) is a remote sensing technology that collects 3-dimensional point clouds of the Earth's surface. This technology is being used for a wide range of applications including high-resolution topographic mapping and 3-dimensional surface modeling as well as infrastructure and biomass studies."[2] The bathymetry data gathered from Lake George is being used to investigate underwater features which may be of use to various models of the lake, including circulation models, food web models, and ecological niche models. In this paper the focus is on ecological niche modeling of invasives in Lake George.

2.2. Invasives in Lake George

Invasive species in Lake George, including the Asian Clam, Zebra Mussels, and Eurasian Watermilfoil, can have a number of negative impacts on the region. They can upset the ecological balance of the lake by outcompeting indigenous species[3], harm water quality through biofouling[4], and hamper recreation in a variety of ways[5]. Vast efforts have been put in place to control the spread of the invasives, with some success. However, such efforts are costly and sometimes ineffective. One of the most effective means of control is to eliminate fledgling colonies of the invasives at new locations before they become well established. The ability to leverage geospatial data in order to predict colonization sites, and then ask questions about them in natural language, would be of benefit to lay people and scientists alike. Our pipeline is one technology that enables this process.

2.3. Extracting knowledge for invasive species prevention

In order to aid with the search for undiscovered invasive sites, ecological niche modeling is performed using the bathymetry data and other maps. The generated predictions are encoded alongside cluster analysis of the data points, as well as data from GeoNames, and the result is RDF conforming to GeoSPARQL standards which can be used to ask interesting questions through simple queries. Because useful analysis is performed ahead of time and put into a convenient format, our scientists working on the Jefferson Project will be able to devote more of their time answering complex questions about the lake instead of handling the data analysis on their own. Section 3.3 explains the specific process of ecological niche modeling.

3. Data Analytics and Integration

3.1. Clustering

Clustering algorithms attempt to group objects such that "a set of objects in the same group are more similar to each other than to those in other groups." [6] There are a number of algorithms which range in their notion of clusters, required a-priori information, etc. In choosing a clustering algorithm, one major consideration is the necessity for a-priori information, specifically the need to specify the number of clusters ahead of time.

In the bathymetry data of Lake George there is an unknown amount of geographical features, so algorithms which do not need a parameter specifying the number of clusters to look for are preferred. Density based clustering techniques are fairly appropriate in this scenario. In this case, a Mean-Shift [7] algorithm is used. The implementation of the algorithm comes from the Scikit-learn package for python. [8] After clustering is performed, convex hulls are generated for each cluster. These convex hulls allow the boundaries of the objects to be represented, thus preserving the structure of the feature, but reduce the amount of points which must be considered for the object. Below is an example of the output format generated by the algorithm, with real data removed due to data restrictions. Currently its intermediate form is plain text, though in the future it will likely be put in some other format like RDF at this stage.

Cluster 0:

$$\begin{array}{lll} X_1 & Y_1 & Z_1 \\ X_2 & Y_2 & Z_2 \\ \vdots & \vdots & \vdots \end{array}$$

Cluster 1:

$$\begin{array}{lll} X_1 & Y_1 & Z_1 \\ \vdots & \vdots & \vdots \end{array}$$

The resulting clusters are combined with maps and other derived products to form a knowledge base of the lake. In order to obtain some of the map data, GeoNames[9] is used.

3.2. GeoNames

"GeoNames contains over 10 million geographical names and consists of over 9 million unique features whereof 2.8 million populated places and 5.5 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes."[9] Further "Users may manually edit, correct and add new names using a user friendly wiki interface." Such flexibility is of benefit to the Jefferson Project, since gaps in data can be filled in. RDF data from the GeoNames API was downloaded for the Lake George region via a simple bounding box query. Figure 1 shows a map of the bounding box, as well as the specific query used to retrieve the results.

This data contains a number of named entities, including islands, bays, rivers, buildings, and much more, that will allow a broader range of questions to be asked about the bathymetry data when the two are combined.

Fig. 1. Overview of the Lake George region courtesy of Google Maps [10], approximating the bounding box used in gathering data from GeoNames. The data from this area was generated with an API call to http://api.geonames.org/

3.3. openModeller

Ecological niche modeling is an analysis which can be uniqely applied to the data obtained from lakes and other natural environments. In previous work, the Environmental Protection Agency has used ecological niche modelling to predict the distribution of invasives in the Laurentian Great Lakes[11]. The tools pro-

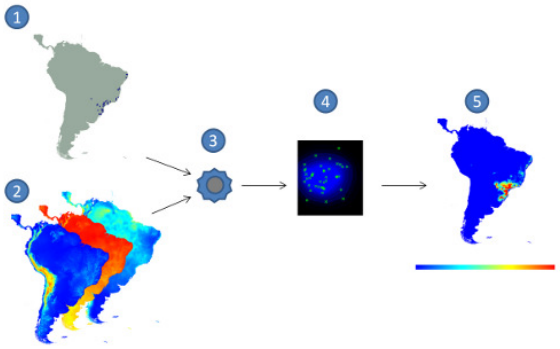

Fig. 2. Correlative approach for Ecological Niche Modelling. (1) species occurrence points, (2) environmental layers, (3) modelling algorithm, (4) model in the environmental space, (5) model projection, with red indicating higher suitability values and blue the lower ones.

[13]

vided in the openModeller package [12] allow for flexible niche modeling to be performed. Input to openModeller is a user defined group of map layers, each of which can represent a different characteristic of the geography at a given latitude and longitude. This allows the user to include as much relevant information as she/he desires. The output is a map of probabilities that a given invasive species will spread to certain areas. Figure 2 shows an illustration of the openModeller workflow.

For the use of openModeller in Lake George, each point on the map had the following data associated with it: depth, distance to nearest boat launch, and existence of at least one invasive (some invasive species are more likely to be found in areas with other invasives). In the future, in order to include even more relevant information, efforts will be made to gather substrate data, nutrient data, temperature data, etc. Currently the Jefferson project platforms which measure such attributes are in the early stages of deployment, though a mass creation of this data is expected soon.

3.4. An example run

With the previous sections in mind, consider the following dataflow. Before any processing is done, there are files containing bathymetric and topological maps of the Lake George region, along with a file containing known locations of certain invasive species in Lake George. Firstly, clustering is performed on the bathymetry and topology. The clusters are stored in plain text in files corresponding one to one with the bathymetry files. Various scripts are then called to translate the bathymetry files into a GeoTIFF format suitable for use with openModeller. Meanwhile another script retrieves GeoNames data through the GeoNames API, using the boundaries of the GeoTIFF files to create a bounding box. The GeoNames data is stored in a file for future use with the other derived products. At this point, other GeoTIFF layers are generated besides the bathymetry. Using the GeoNames entities, distances to the nearest boat launch or marina are calculated for each point of the bathymetry. Similarly, in another layer, each point is designated as either containing an invasive or not using a boolean. All layers are stored as separate files. openModeller is then called to run on all layers, producing a single map of probabilities. This map of probabilities is examined point by point, with points over a certain threshold being labeled as at risk for the given invasive. These points are then mapped to their corresponding clusters, and the cluster is tagged as at risk. Now that all clusters are labeled, they are ready to be described using RDF.

4. Knowledge Organization

4.1. GeoSPARQL

With analysis having been performed, the derived products are ready to be described in GeoSPARQL conformant RDF. All GeoSPARQL triples generated are based on the features of the lake and the surrounding area, i.e the clusters generated and nodes extracted from GeoNames. Thus, in modeling the spread of invasive species in Lake George, the simulation is run and the results are simply added as triples describing the features of the lake. Each feature is described with a combination of concepts from the Jefferson Project's ontology, RDF, and the OpenGIS ontology. Clusters are serialized into RDF format by the python library RDFlib [14].

When possible, a feature's geometry is described as a polygon using the OpenGIS simple features ontology, with lats and longs using the WGS84 projection. However, not all nodes extracted from GeoNames data have defined boundaries, so in this case the point concept from Open GIS is used. Triples which do not correspond to geometries can describe a range of knowledge, from labels extracted from GeoNames data, to the ecological niche modeling analysis. For example, a cluster or GeoNames node which has been identified as a suitable location for an invasive species by openModeller is marked as atRiskForInvasive: Inva-

siveSpeciesX. Listing 1 shows GeoSPARQL conformant triples in n3 format representing one of the clusters of the lake. Identifying points have been removed due to data restrictions.

```
jp_bath:Cluster10004 a geo:Feature ;
    jp_bath_ont:atRiskForInvasive
    "Asian clam"@en ;
    dbpedia_owl:depth 9.03e+01,
        9.06658e+01,
        9.133e+01 ;
    geo:hasGeometry
    jp_bath:Cluster10004Poly ;
    foaf:name "Cluster 10004"@en .
```

Listing 1: Sample GeoSPARQL

Using RDF to describe the results of analysis is valuable not only because of it's suitability for question answering, but also because it is human readable. It is to both the scientists' and the informaticists' benefit to be dealing with information in an easy to understand format, as opposed to large amounts of numerical data. Validation of conclusions becomes easier since the sources are more easily understandable. Analysis is clear, and placed in a meaningful context alongside named entities and objects which scientists recognize.

4.2. The need for clusters

Subdividing the lake and surrounding region into smaller features using clustering allows for a description of the lake based on human-level concepts, eliminating the need to consider individual data points. Since points in a bathymetric cluster are guaranteed to be similar to one another in depth and location, labeling a cluster as at risk for invasive species based on the existence of an at risk point is a valid operation, whereas using map data alone would prevent such a practice. For example, assume an area of a bay in the lake was classified as at risk for an ivasive species. Labeling the entire bay as at risk for that species could not be done, since the bathymetry of the bay may change significantly from location to location. Since GeoSPARQL supports topological reasoning, including the DE-9IM,[15] queries can be constructed looking for clusters within the named regions. This design allows information about more general named regions to be deduced through the clusters it contains, while maintaining an accurate representation of the performed analysis.

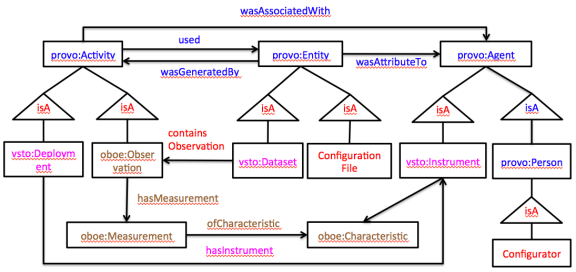

Fig. 3. HAS-Net O leverages PROV, OBOE, and VTSO in order to describe the sensors and instruments themselves, and related events such as instrument deployment. This will maintain structured metadata about the sensors and instruments, providing valuable context behind the measurements gathered during the course of the project.

4.3. HASNetO

The Jefferson Project's use of ontologies facilitates the integration of information coming from multiple sources, as described in our use cases, and is thus easily reusable and combinable [16]. We use the Human-Aware Sensor Network Ontology (HASNet-O) [17], which is an orchestrated integration and extension of three well-established ontologies: the Extensible Ontology for Observations (OBOE) [18], the Virtual Solar-Terrestrial Observatory's ontology (VSTO) [19], and the World-Wide Web's Provenance Language (PROV) [20]. HASNetO's name is derived from the fact that important metadata, namely information about the human activities involved in setting up a sensor network, are often ignored or poorly documented at best. HASNetO provides a vocabulary that enables scientists to encode metadata that describes how humans calibrate instruments, how humans participate in the process of deployment instruments, and how humans may affect tthe quality of observation data by changing the instruments' parameters. Figure 3 shows an outline of the HASNetO ontology.

OBOE provides HASNetO with concepts to describe entities and their observable characteristics. Characteristics are entity's properties that are measurable. For example, entities of interest like 'Air' cannot be directly measured, while air's characteristics like temperature and humidity can. OBOE also defines that an observation is a collection of measurements of a characteristic where all the measurements of an observation were performed under common conditions, i.e the observation context.

VSTO provides HASNetO with concepts to describe the infrastructure of sensor networks. A sensor network infrastructure is composed of platforms that

host instruments that in turn host detectors. It should be noted that VSTO does not have a 'Sensor' concept, and often the term sensor is used to refer to detectors, instruments, or combinations of the two. Also of note is that it may be challenging to identify detectors in environments less sophisticated than that of the Jefferson Project. For example, an ordinary thermometer, which is an instrument, is often presented as a single gadget that has an internal, non-detachable detector. VSTO also defines that a deployment is an activity of installing an instrument/detector combination at a given platform, and enabling the instrument/detector combination to perform observations.

Lastly, PROV provides HASNetO with high-level concepts that enable HASNetO to fully document the entire spectrum of sensor network activities, including both human activities and machine (automated) activities. For example, PROV enables HASNetO to see both VSTO's deployments and OBOE's observations as PROV activities, where observations are subactivities within deployments. Moreover, the entire lifecycle of deployments and observations can be documents including the identification of agents involved in the process of executing these activities, their roles, and any information derived from the activities.

In short, HASNetO allows important provenance and metadata to be available alongside the data and derived products. Further, such metadata is stored in query-able triples. Storing metadata like this facilitates validation of scientific conclusions which may be derived from data and the derived products.

5. Analysis

5.1. Sample Queries

As previously mentioned, early detection of invasive colonies is the best means of controlling an invasive's spread. The Lake George Association performs yearly inspection of varoius sites around the lake looking for Asian Clams and other invasives. Knowing what at risk sites are nearby to a given team's location during this search would be of great benefit. The query in listing 2 shows how to easily access this information.

```
SELECT *
{?s spatial:withinCircle
 (43.44 -73.67 .25 'miles');
 gn:name ?name ;
 jp_bath_ont:atRiskForInvasive
 ?invasive }
```

Listing 2: Nearby Invasives Query

```
SELECT *
{?s dbpedia_owl:depth ?depth;
 gn:name ?name ;
 jp_bath_ont:atRiskForInvasive
 ?invasive }
ORDER BY DESC(?depth)
LIMIT 1
```

Listing 3: Maximum At Risk Depth Query

Also of interest are the environments in which the invasives are predicted to live. A scientist can examine various aspects of at risk clusters, such as depth, which listing 3 demonstrates. The query will return the maximum depth of a cluster considered at risk for an invasive.

Currently the obvious limitation of the triples is the variety of knowledge which they encode. Queries are limited to asking questions about invasives, and various physical attributes of the clusters. More analysis will be put into triples soon.

5.2. Knowledge to Data Ratio

There are a total of 642,783 triples (not including triples from Geonames) generated via the outlined process, in comparison to 108,277,419 data points from the original files. This is a compression rate of about 168 points per triple, which is a fairly significant reduction. The advantages inherent to such a reduction are obvious, such as lowering the required storage space and increasing the speed of processing.

6. Future Directions and Conclusion

6.1. Computer Vision Application

While cluster analysis of the bathymetry data is useful, it may also be useful to identify standalone topo-

logical features in the bathymetry, such as channels, ravines, etc. To automatically identify such features, computer vision techniques would need to be applied.

6.2. Knowledge Extraction from Texts

Natural language processing and knowledge extraction could make siginificant contributions to the approach already being taken. For example, if the existing sites of certain invasive species in Lake George could be determined from reports released by interested parties, the process of ecological niche modeling could be automated even further. Searching for texts about the entities named in the GeoSPARQL triple store could allow new knowledge to be added automatically, if the knowledge could be extracted from those texts.

6.3. Using Watson

Another way of interacting with natural language data is to use a cognitive computing platform like IBM's Watson. Since Watson runs on both Linked Data and data in the form of natural language texts, the existing GeoSPARQL could be combined with texts about Lake George, Hydrology, Biology, etc, to serve as a knowledge base for Watson. An obvious advantage of leveraging Watson's ability to handle natural langauge is the increase in human analysis which becomes available to the researchers of the project. Not only do decades of research papers become available to researchers, but also real time texts, e.g newspapers which may describe happenings in the lake faster than any other form of data. Rensselaer Polytechnic is in a unique position to leverage this technology because of its relationship with IBM, its access to the IBM Watson academic software, and its long standing usage of Watson and similar technologies.

6.4. Conclusion

In summary, the outlined methodology provides a method of analyzing geospatial data and placing the results in linked data, suitable for use by QA systems. Clustering algorithms are used to split spatial data into regions, which are combined with named entities to create a context for the regions which is easily digestible by humans. Ecological niche modeling is performed using geospatial data, and the analysis is placed in the context of regions and named entities. The knowledge extraction and formatting pipeline was described. The motivations of the Jefferson Project were described, along with how the approach taken is helping researchers combat invasive species. Next the data analyses employed were outlined, as were the methods used to transform these analyses into geospatial knowledge encoded as linked data. The use of GeoSPARQL was described, along with its uses. The compression achieved by transitioning from the data's raw form to RDF was highlighted.

7. Acknowledgements

The authors wish to recognize various parties which have made an impact on this research: The Darrin Freshwater Institute, specifically Rick Relyea and Jeremy Farrell, International Business Machines, specifically Eli Dow, Mike Passow, Michael Kelly, Harry Kolar, and Mike Henderson, and Rensselaer Polytechnic institute, specifically James Hendler, Katie Chastain, and Laura Kinkead.[1]

References

[1] Rensselaer Polytechnic Institute. New project aims to make new york's lake goerge the smartest lake in the world. 06 2013. [Online; accessed 4-May-2014].

[2] Wikipedia. Light detection and ranging (lidar), January 2012. [Online; accessed 28-December-2014].

[3] John D Madsen, JW Sutherland, JA Bloomfield, LW Eichler, CW Boylen, et al. The decline of native vegetation under dense eurasian watermilfoil canopies. *Journal of Aquatic Plant Management*, 29:94–99, 1991.

[4] Ronaldo Sousa, Antonio JA Nogueira, Miguel B Gaspar, Carlos Antunes, and Lúcia Guilhermino. Growth and extremely high production of the non-indigenous invasive species< i> corbicula fluminea</i>(müller, 1774): Possible implications for ecosystem functioning. *Estuarine, Coastal and Shelf Science*, 80(2):289–295, 2008.

[5] Wikipedia. Myriophyllum spicatum — Wikipedia, the free encyclopedia, 2014. [Online; accessed 11-November-2014].

[6] Wikipedia. Cluster analysis — Wikipedia, the free encyclopedia, 2014. [Online; accessed 4-May-2014].

[7] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] GeoNames. About geonames, 2014. [Online; accessed 11-November-2014].

[10] Google, 2014. [Online; accessed 27-December-2014].

[11] U.S Environmental Protection Agency. Predicting future introductions of nonindigenous species to the great lakes. 2008.

[12] Mauro Enrique de Souza Muñoz, Renato De Giovanni, Marinez Ferreira de Siqueira, Tim Sutton, Peter Brewer, Ricardo Scachetti Pereira, Dora Ann Lange Canhos, and Vanderlei Perez Canhos. openmodeller: a generic approach to speciesâĂŹ potential distribution modelling. *GeoInformatica*, 15(1):111–135, 2011.

[13] openModeller.

[14] RDFlib. rdflib documentation, 2014. [Online; accessed 10-November-2014].

[15] David Abel and Beng Chin Ooi. *Advances in Spatial Databases: Third International Symposium, SSD'93, Singapore, June 23-25, 1993. Proceedings*, volume 692. Springer, 1993.

[16] D.L McGuinness, Paulo Pinheiro, E.W Patton, and K Chastain. Semantic escience for ecosystem understanding and monitoring: The jefferson project case study. Proceedings of AGU Fall Meeting 2014, Moscone Center, San Francisco, CA, US, December 2014.

[17] Paulo Pinheiro and D.L McGuinness. Provenance-enabled integration of sensor network data. Proceedings of AGU Fall Meeting 2014, Moscone Center, San Francisco, CA, US, December 2014.

[18] Joshua Madin, Shawn Bowers, Mark Schildhauer, Sergeui Krivov, Deana Pennington, and Ferdinando Villa. An ontology for describing and synthesizing ecological observation data.

[19] Deborah McGuinness, Peter Fox, Luca Cinquini, Patrick West, Jose Garcia, James L Benedict, and Don Middleton. The virtual solar-terrestrial observatory: A deployed semantic web application case study for scientific research. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 22, page 1730. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[20] T Lebo, S Sahoo, D McGuinness, K Belhajjame, J Cheney, D corsar, D Garijo, S Soiland-Reyes, S Zednik, and J Zhao. Prov-o: The prov ontology, w3c recommendation. *World Wide Web Consortium*, 2013.