Geographic Literacy in Text Analytics: Developing and Applying OpenSextant

Marc Ubaldino
ubaldino @ mitre.org
Co-founder OpenSextant.org
The MITRE Corporation

The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

Jan-30 2020 ABCD GIS Seminar Series
hosted by Center for Geographic Analysis
Harvard University

Introducing OpenSextant

http://opensextant.github.io/

- Open Source, since 2013
- Geo/Temporal Info Extraction
- Government-supported, MITRE-operated
  - More info: http://www.mitre.org/
- Focus: unstructured data & text processing

The OpenSextant Project

A geospatial stuff project

Info Extraction Projects

OpenSextantToolbox - A geotagger and entity extractor employing GATE

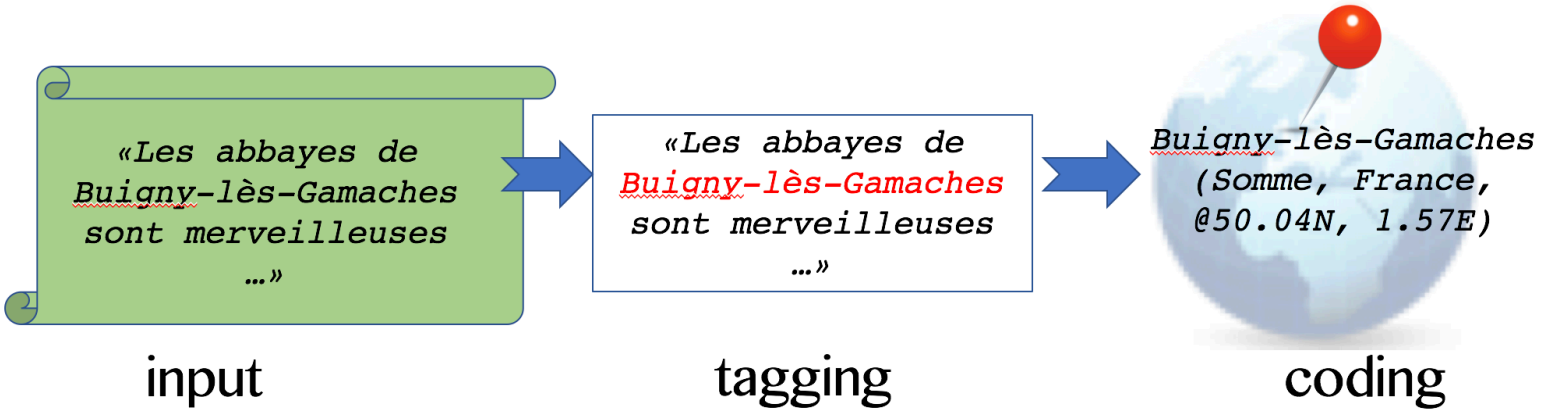
Xponents - Geotagging APIs to work with gazetteers and multilingual information extraction for geography, date/time, patterns. This suite makes use of Tika for rendering inputs and GISCore (below) for output formats.

SolrTextTagger - A text tagger based on Lucene/Solr. (As of Solr 7.4 this tagger tool is a formal request handler in Solr)

Gazetteer - Pipeline project to render world-wide "geo names" data into gazetteers used by these projects

MITRE

The General Problem



OpenSextant provides open source libraries and methodologies for condition input text, *geotagging* and *geocoding* that text.

Additionally, these tools offer other related natural language processing (NLP) for date/time extraction, keyword and named entity extraction.

The topic of this presentation to expose the challenges in developing and evolving these capabilities for addressing world language and geography in a practical manner for various use cases and users.

 Approved for Public Release; Distribution Unlimited, Case# 20-0159

MITRE

Agenda

Purpose for today: raise awareness, develop community, seek R&D collaboration around OpenSextant.

Exhibit recent activities and describe future direction. Present **challenges** and **technical solutions**.

Discuss *Geographic Literacy* in software and people throughout.

Approved for Public Release; Distribution Unlimited, Case# 20-0159

Geographic Literacy

My working definition:

Geographic literacy (GL) is the ability to understand and the depth of understanding of geography with respect to language, culture, history, and context.

Major components of this include:

- Thoroughness of reference knowledge available
- Ability to recognize geographic references
- Application of common sense when interpreting *found* geography in data

GL in Software:

How do geographic functions vary when developing library, desktop app, mobile app, cloud or AI solutions?

What are the expectations of developer vs. end user?

GL in People:

What is expected of students, public officials, civilians, mobile users regarding use of maps, directions, and obscure geography in daily information?

Approved for Public Release; Distribution Unlimited, Case# 20-0159

The Opportunity

GL in Software:

OpenSextant shows MITRE has a good handle on the technical aspects, i.e., how to research and build solutions.

GL in People:

OpenSextant and other works show MITRE has a good handle on the issues within a government audience.

Can OpenSextant be applied to an educational audience to aid teachers and students to navigate geographic-bearing information in texts and data sets of all sizes? Educational audience could be a classroom of K-12, college, or professional, for example.

This is good open source community work that should reach other audiences.

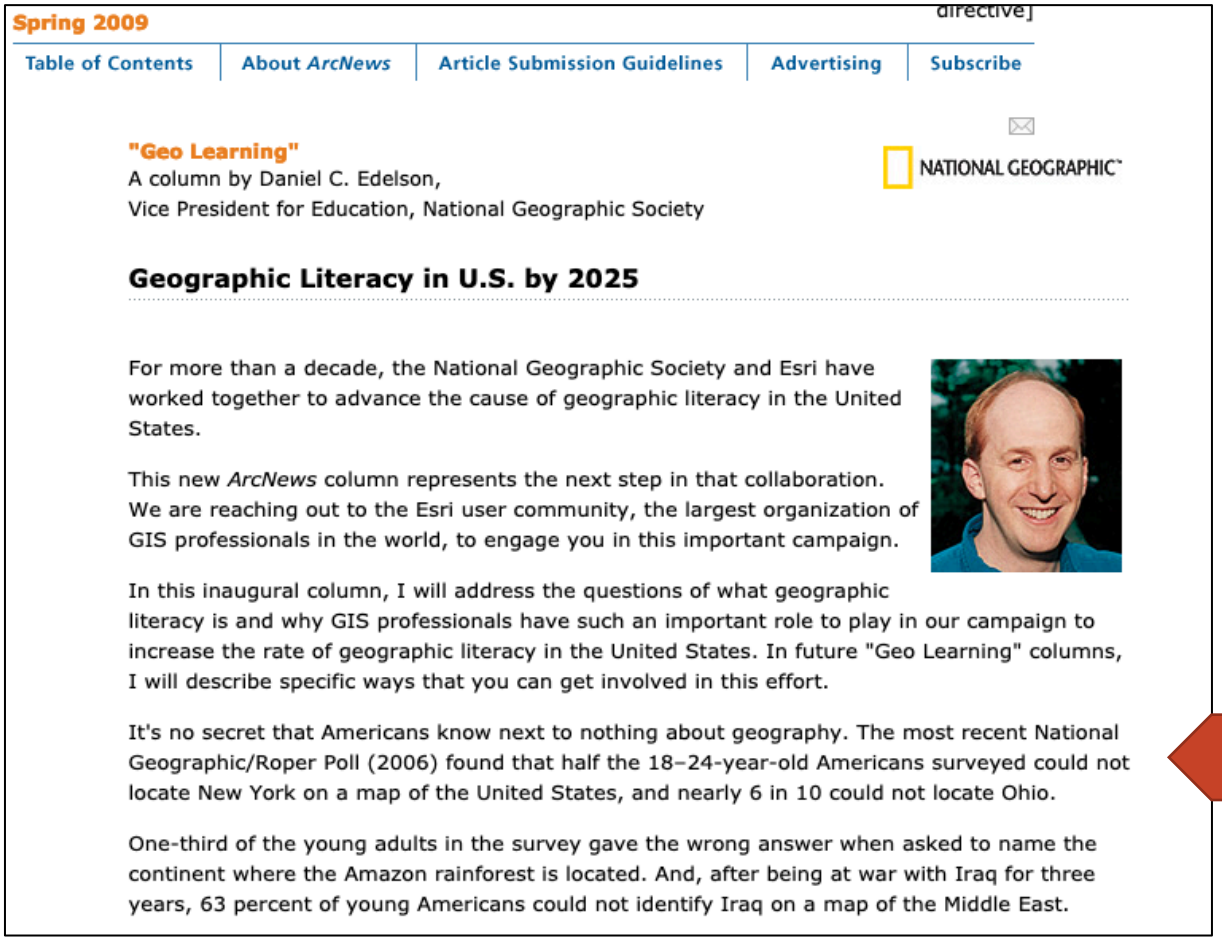
Approved for Public Release; Distribution Unlimited, Case# 20-0159

Exemplars and Pioneers of GL

National Geographic: Educational Standards for Geographic Literacy quantitatively defined
https://www.nationalgeographic.org/standards/national-geography-standards/

American Association of Geographers
http://www.aag.org/geocapabilities and many other educational efforts

Esri "Geo Learning" column. Published 2009. Accessed 2019.
https://www.esri.com/news/arcnews/spring09articles/geographic-literacy.html.



Spring 2009

directive]

Table of Contents | About *ArcNews* | Article Submission Guidelines | Advertising | Subscribe

"Geo Learning"
A column by Daniel C. Edelson,
Vice President for Education, National Geographic Society

NATIONAL GEOGRAPHIC

Geographic Literacy in U.S. by 2025

For more than a decade, the National Geographic Society and Esri have worked together to advance the cause of geographic literacy in the United States.

This new *ArcNews* column represents the next step in that collaboration. We are reaching out to the Esri user community, the largest organization of GIS professionals in the world, to engage you in this important campaign.

In this inaugural column, I will address the questions of what geographic literacy is and why GIS professionals have such an important role to play in our campaign to increase the rate of geographic literacy in the United States. In future "Geo Learning" columns, I will describe specific ways that you can get involved in this effort.

It's no secret that Americans know next to nothing about geography. The most recent National Geographic/Roper Poll (2006) found that half the 18–24-year-old Americans surveyed could not locate New York on a map of the United States, and nearly 6 in 10 could not locate Ohio.

One-third of the young adults in the survey gave the wrong answer when asked to name the continent where the Amazon rainforest is located. And, after being at war with Iraq for three years, 63 percent of young Americans could not identify Iraq on a map of the Middle East.

... GL as it relates to **people**

c.2009. And today?

MITRE

... and GL as it relates to **software** is dominated by standards bodies and GIS capability communities, including:

- **OSGeo.org** (https://www.osgeo.org/): enterprise grade GIS ecosystems
- **LocationTech @ Eclipse** (https://projects.eclipse.org/projects/locationtech): Big-data spatial analytics, libraries and location-aware utilities
- **Open Geospatial Consortium** (http://opengeospatial.org). GIS standards.
- **ISO, FIPS**, etc – information standards for world geography
- Etc.

However, these groups do not touch on the harder topics of interpreting geographic entities in unstructured data or text – this is still an art form with little convergence on best practices (in my opinion).

Our Focus

Substantial effort put into evaluating available open source techniques, mainly on social media:

- "A Practical Guide for the Effective Evaluation of Twitter User Geolocation", Mourad, et al 2019.
- "A pragmatic guide to geoparsing evaluation Toponyms, Named Entity Recognition and pragmatics", M. Gritta, et al. 2019.

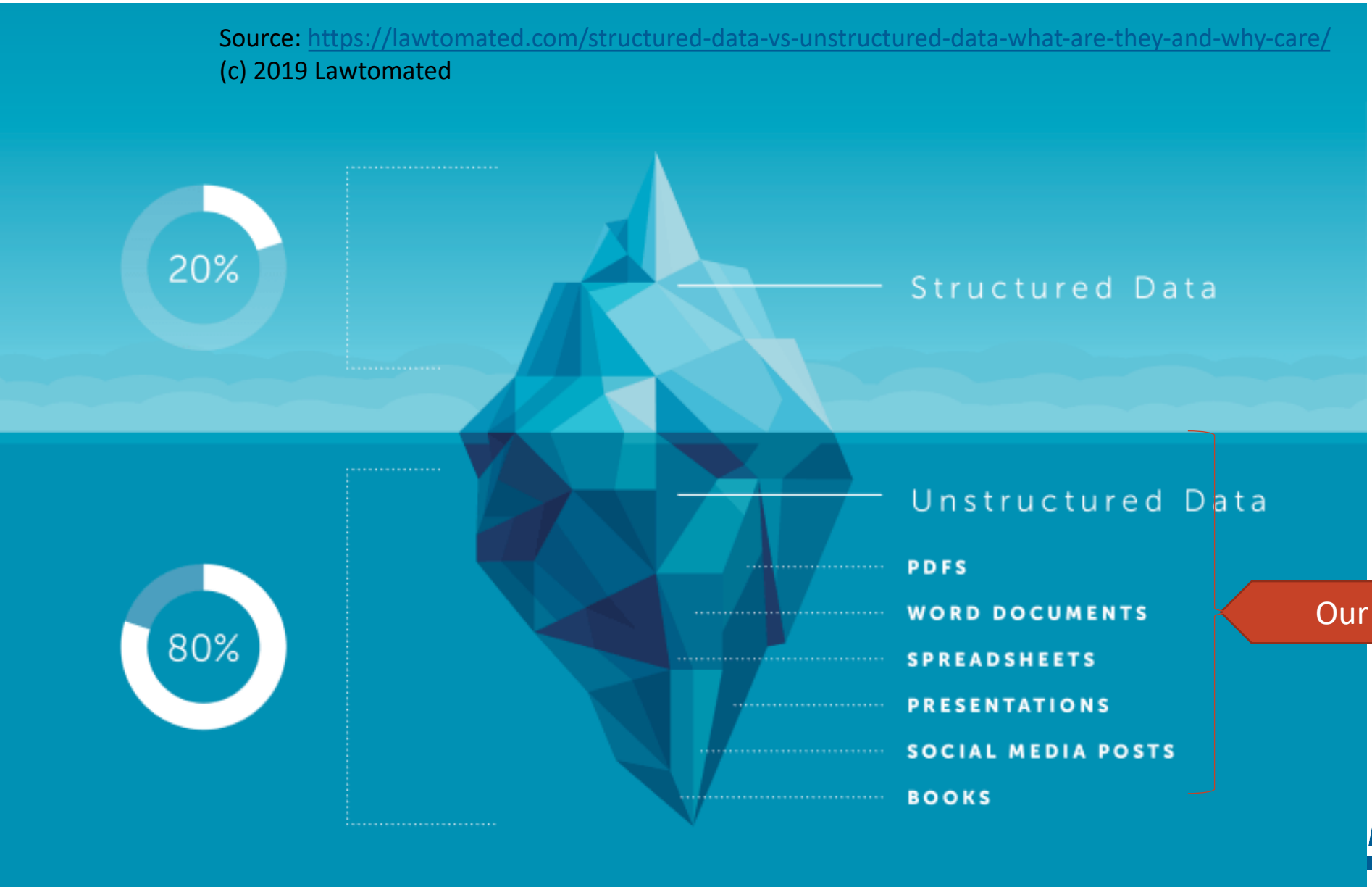
Approved for Public Release; Distribution Unlimited, Case# 20-0159

Why Does it matter?

Many sources (including IDC & Gartner), estimate an 80/20 ratio of unstructured to structured data, primarily within enterprises. No corresponding ratio for public domain data known.

Unstructured data **can** result in "geospatial intelligence" only after substantial amount of work (*art & science*) is applied to it.

Source: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/
(c) 2019 Lawtomated



MITRE

Addressing Technical Challenges in GL with OpenSextant

Challenges

0. Driving Factors
1. Geotagging and Geocoding
2. Working in Any Language
3. Adding, Extending and Maintaining Gazetteers and Reference Data
4. Delivering the Capability

Approved for Public Release; Distribution Unlimited, Case# 20-0159

Driving Factors

Unstructured Data processing is a "wild-west" – be ready for anything and loose general assumptions about data/text

Volume of reference data is large

Quality and recency of data may be tied to connectedness of a solution

Localization (language, colloquialisms, culture)

Speed

Cost of on-site solutions, cost of R&D, etc

18 million place names for 10 million locations...

Connectivity is assumed

Half of the world speaks 6 languages.

Approved for Public Release; Distribution Unlimited, Case# 20-0159

Building Geographically Literate Solutions

Objectives

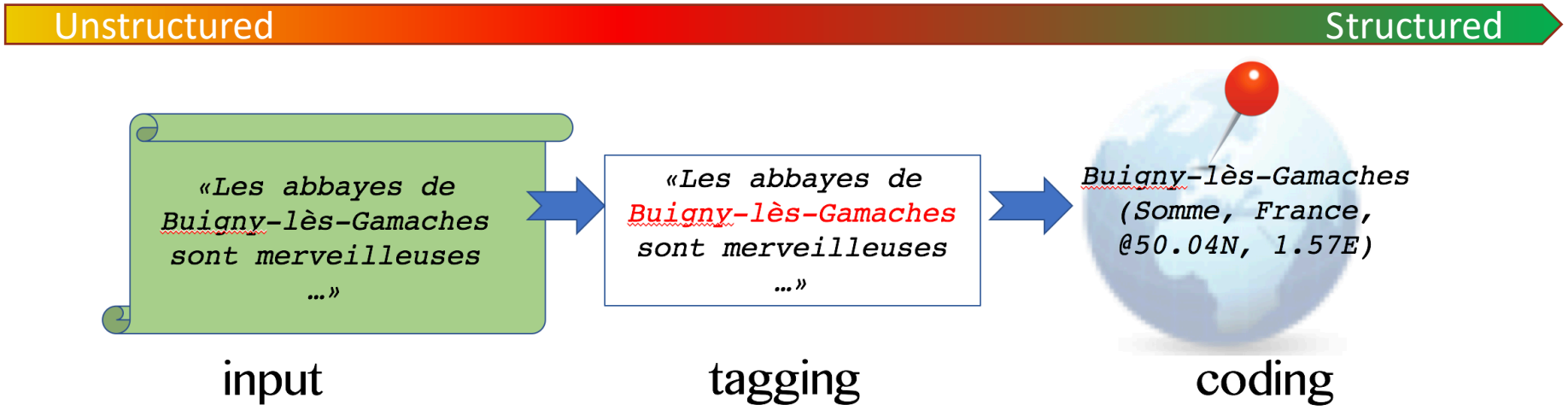
- ❑ **Derive potential "location intelligence"**
- ❑ **Globally aware, Adaptable to local focus**
- ❑ **Provide means for practical geolocation, that is justifiable and intuitive**

Compromises and Constraints

- ❑ **Balance "good enough" with thoroughness**
- ❑ **Compactness and portability**
- ❑ **Operate offline in the field or other Intranet contexts**
- ❑ **Speed and Interactivity depends on use case**
- ❑ **Modular design**
- ❑ **Maintain attribution to source data**

Approved for Public Release; Distribution Unlimited, Case# 20-0159

Challenge #1. Geotagging & Geocoding



- **What language and format is the data?**
- **Is genre and size of data relevant?**

- **What geography matters to you?**
- **What entities are important?**
- **What reference data do you trust?**
- **How does the NLP treat the input text?**

- **What geographic/geospatial metadata matters most?**
- **What is the intended use of "point" or "features"?**
- **How do you link coded output to original input – in map, database, etc?**
- **How do you convey this "inferencing" properly vs. sufficiently?**

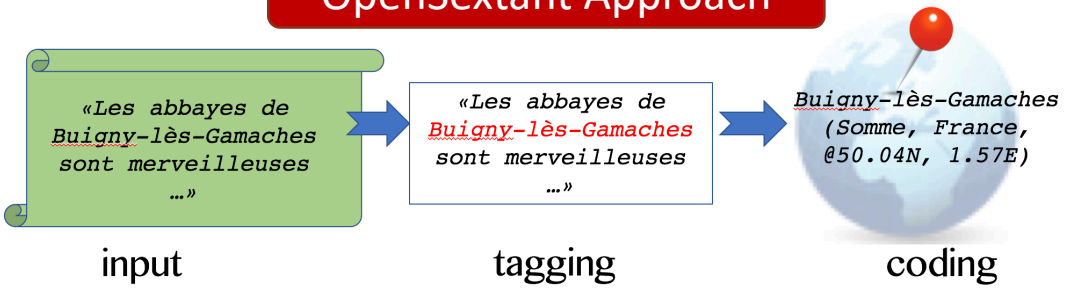
Example scenarios:

French vs. Japanese text
All-Upper Case text
Social vs. Journalistic genre
Extracting national parks vs. towns

MITRE

Challenge #1. Geotagging & Geocoding

OpenSextant Approach



- ❑ **Acquire text, detect language and character encoding**
- ❑ **Determine text case**
- ❑ **Detect geo coordinates**
- ❑ **Detect potential named places, countries, and nationalities**
- ❑ **Identify easily knowable non-places from stop words or related lexica**
- ❑ **Resolve hierarchical relationships between locations and boundaries**
- ❑ **Weight locations against name tags**
- ❑ **Choose location, providing evidence and confidence**

See: https://opensextant.github.io/Xponents/

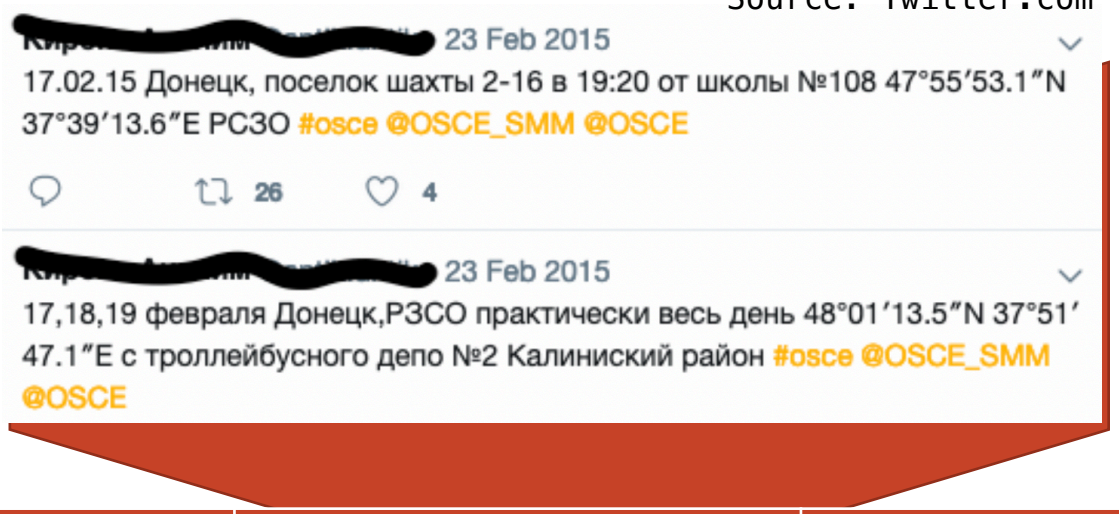
Data	ISO-639 Language Codes	ISO-3166 country codes, nationalities, Geonames, stop words by language, Well-known Entities
Tools	Tika, LangDetect	Solr, Lucene FST
Open Sextant	Text Utilities	Gazetteer, XCoord, PlaceGeocoder, *"Geocoder Handbook"*

MITRE

Example

Source: Twitter.com

input

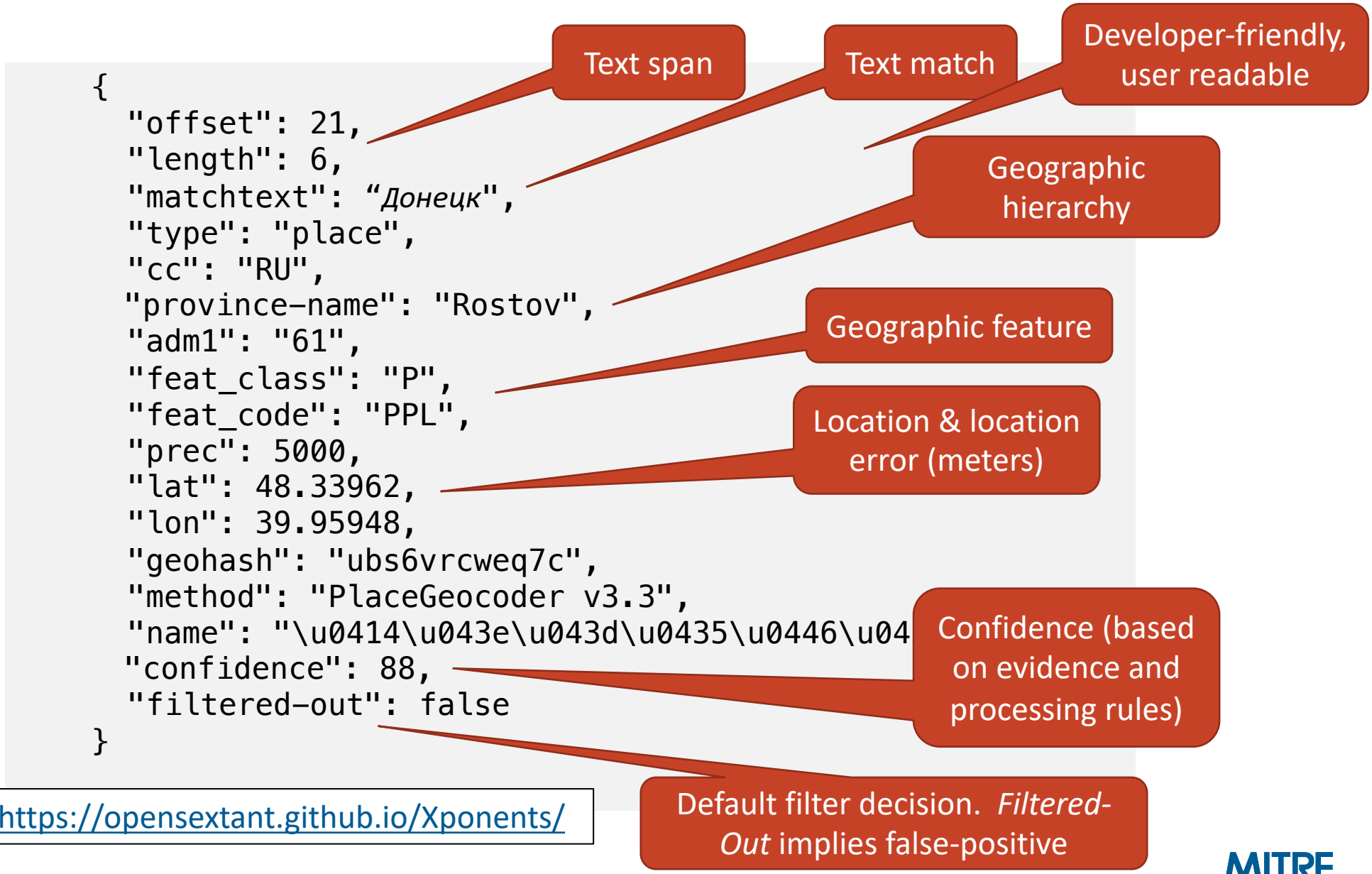


output

Found Entities	Type	Coding	OpenSextant Method
47°55′53.1″N 37°39′13.6″E	Coordinate	+47.93142, +37.65378, UA.05	XCoord (regular expressions), Gazetteer reverse lookup
48°01′13.5″N 37°51′47.1″E	Coordinate	+48.02042, +37.8631, UA.05	
Донецк	City name	Populated place @48.339, 39.956 in Rostov, RU (RU.61)	Entity tagging (Gazetteer w/Solr FST), Rules, Filters
23 Feb 2015	Date	2015-02-23	XTemp (regular expressions)
OSCE	Organization	“Org.Organization for Security and Co operation in Europe”	Entity tagging (”JRC” lexicon w/Solr FST)

MITRE

OpenSextant "Xponents" Schema for Geotagging and Geocoding



See: https://opensextant.github.io/Xponents/

MITRE

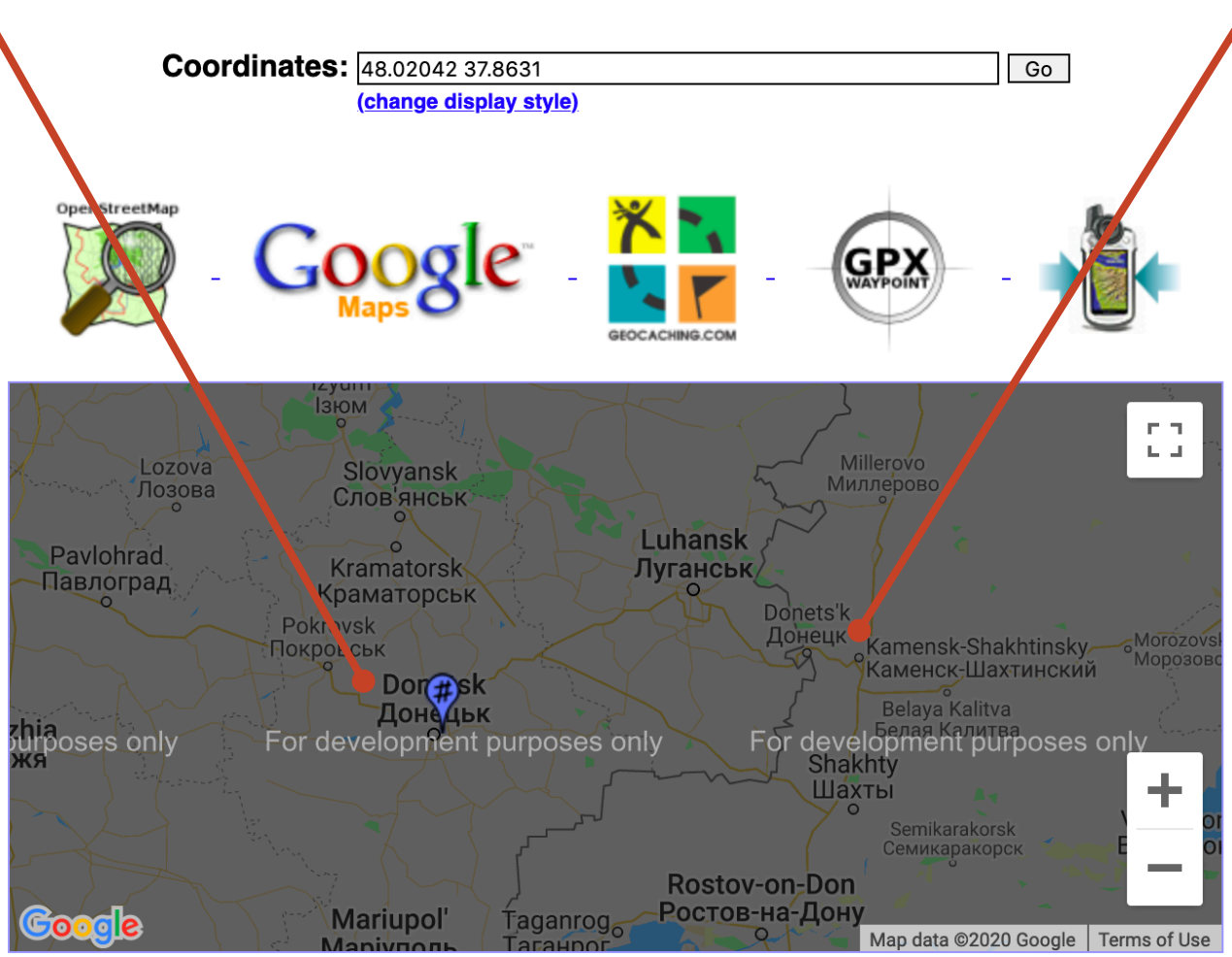
Visual Inspection

Both Locations / Geocodings are relevant based on Tweet, however are in different Countries and about 50KM apart.

Testing & validating foreign language examples is detective work!

Донецьк *= "Donetsk", city in Ukraine, containing coordinates in tweet*

Донецк *= "Donets'k" in tweet.*



Source: http://geohash.org/ubs6v /

MITRE

Challenge #2. Working in Any Language

Observation: Developing decent geotagging performance in any language requires an army of dedicated linguist/geographers needed to develop some ground truth

- to customize and test rules or
- to train language models (LM)
- or both.

Until you find that army of colleagues, some resources are helpful:

- **MediaEval** (http://www.multimediaeval.org/) orchestrates evaluation tasks and data set curation
- **Joint Research Centre** (https://ec.europa.eu/jrc/en/language-technologies) supports relevant EU research, offering multilingual data sets
- **Apache Lucene** (http://lucene.apache.org/) language processing resources and techniques
- Many more....

Approved for Public Release; Distribution Unlimited, Case# 20-0159

Challenge #2. Working in Any Language

OpenSextant Xponents Approach

Objective: Support widest-range of languages and scripts.

Major world-language groups supported:
1. European,
2. Chinese/Japanese/Korean (CJK),
3. Arabic

Static resources

OpenSextant Solr Gazetteer

Key Fields
Name
Name_tag
Name_AR
Name_tag_AR
Name_CJK
Name_tag_CJK

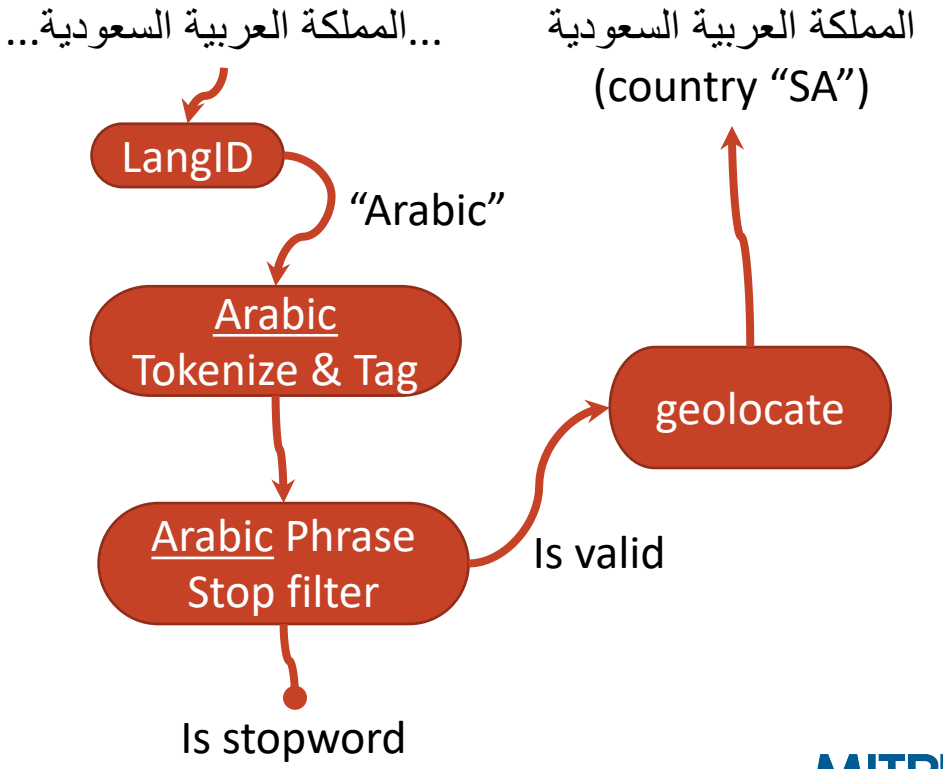
3.5 GB in size

JRCNames, etc. well-known entities (10 langs)

LangDetect Models (50+ langs)

Lucene and custom Stop word lists (35 langs)

Processing Flow



Approved for Public Release; Distribution Unlimited, Case# 20-0159

MITRE

Challenge #3. Adding, Extending and Maintaining Gazetteers and Reference Data

Multiple sources of reference data, compounded by the need to adapt and extend that data with additional insights to support geotagging.

Among Largest data sets are USGS Domestic Names and NGA Geonames, but Geonames.org provides the most diverse with almost 400 discrete sources covering the world. https://www.geonames.org/datasources/

- What sources do you trust?
- What are official sources vs. organic, crowd-sourced, localized or ad hoc?
- How often do you update?
- How can you insert your own gazetteer data?
- Can you afford to pay for such a service in R&D?

Approved for Public Release; Distribution Unlimited, Case# 20-0159

Challenge #3. Adding, Extending and Maintaining Gazetteers and Reference Data

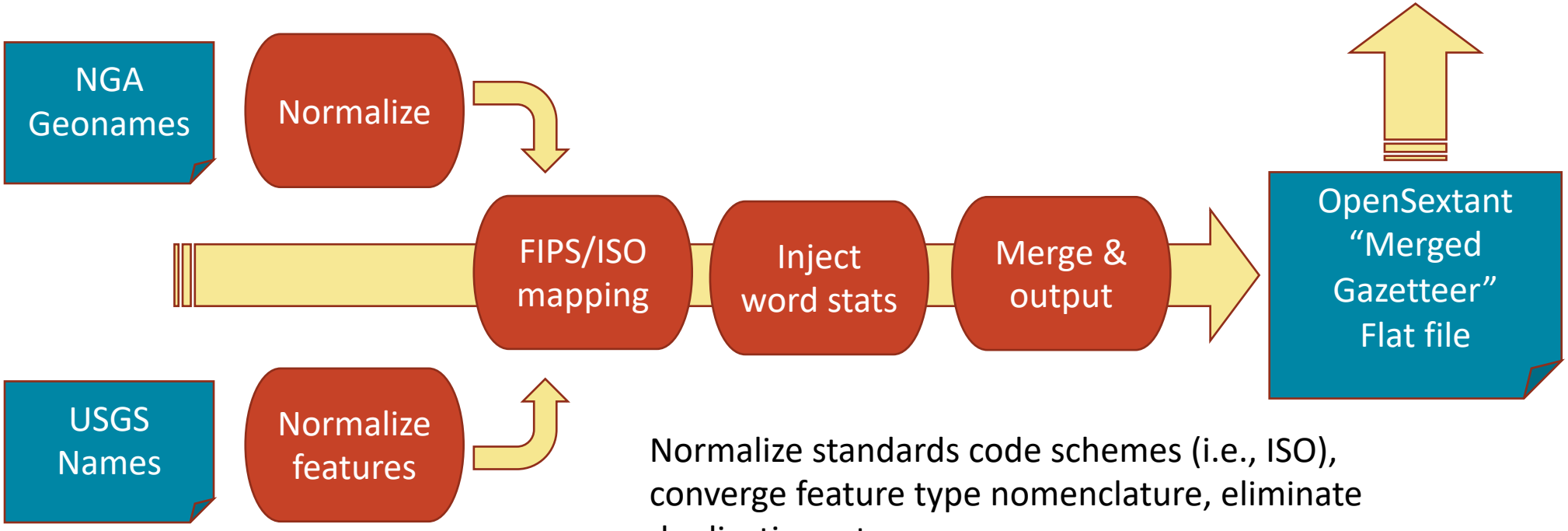
OpenSextant Approach

http://opensextant.github.io/Gazetteer/

Step 1. Generate a Merged, Global Gazetteer

Step 2. Index w/Solr >>

OpenSextant Solr Gazetteer



Normalize standards code schemes (i.e., ISO), converge feature type nomenclature, eliminate duplication, etc.

Approved for Public Release; Distribution Unlimited, Case# 20-0159

MITRE

Challenge #4. Capability Delivery

What form do end user applications take?

Where do they need to use them?

What training and guidance do they need on interpreting “inferred geography”?

analyst

Analyst support team

Developers/Integrators

World-studies classroom

Geography or Data scientist lead

Approved for Public Release; Distribution Unlimited, Case# 20-0159

MITRE

Challenge #4. Capability Delivery

OpenSextant Xponents Approach

Analyst teams benefit:

- ❑ Ability to geotag offline, on laptop
- ❑ Fast (50KB/sec per CPU)
- ❑ Compact (4GB install)
- ❑ Desktop sample app can crawl folders and outputs to GIS formats without any services or network!

analyst

Analyst support team

Geography or Data scientist lead

Developers/Integrators

Developers, Integrators & Data Scientists consult GitHub site https://opensextant.github.io/Xponents/ which provides overview of project including:

- ❑ **Maven**-published (org.opensextant) APIs offer base functions
- ❑ **Docker** image offers complete build of **webservice** & **gazetteer** *https://hub.docker.com/r/mubaldino/opensextant*
- ❑ Java & Python client for webservice
- ❑ Methodology in "**Geocoder Handbook**", https://opensextant.github.io/Xponents/doc/Geocoder_Handbook.html

World-studies classroom

MITRE

Overall Performance

Xponents v3.3 Measured against 400 ACE* (c.2006) conference documents.

Geotagging: Identifying places:

- Recall = 0.97
- Precision = 0.835
- F1 score = 0.895

Geocoding: Correct geolocation of localities (within 1km): 85%

In 2013 MITRE created a “geocoder evaluation” task with several datasets. This activity has not yet been open sourced. “Tagging” ground truth exists, however not for “Geolocation”.

* ACE: https://www.ldc.upenn.edu/collaborations/past-projects/ace

Example Applications

Enrichment Pipeline and Document View: News scrapping and geographic review.

Source: https://github.com/OpenSextant/Xponents/blob/master/doc/LuceneRevolution17-Xponents%2C14Sept2017.pdf

Current Events: Mexico

Title: The Latest: Mexico quake: Hotel collapses in Oaxaca
Origin: Boston Herald
Published: 2017-09-08
Original(s): Original Item Online

....

The governor of the Mexican state of Chiapas says that at least three people have been killed in his region in a massive earthquake that hit off the country's coast.

Gov. Manuel Velasco told Milenio TV that the deaths occurred in San Cristobal de las Casas. He also said that the quake damaged hospitals and schools.

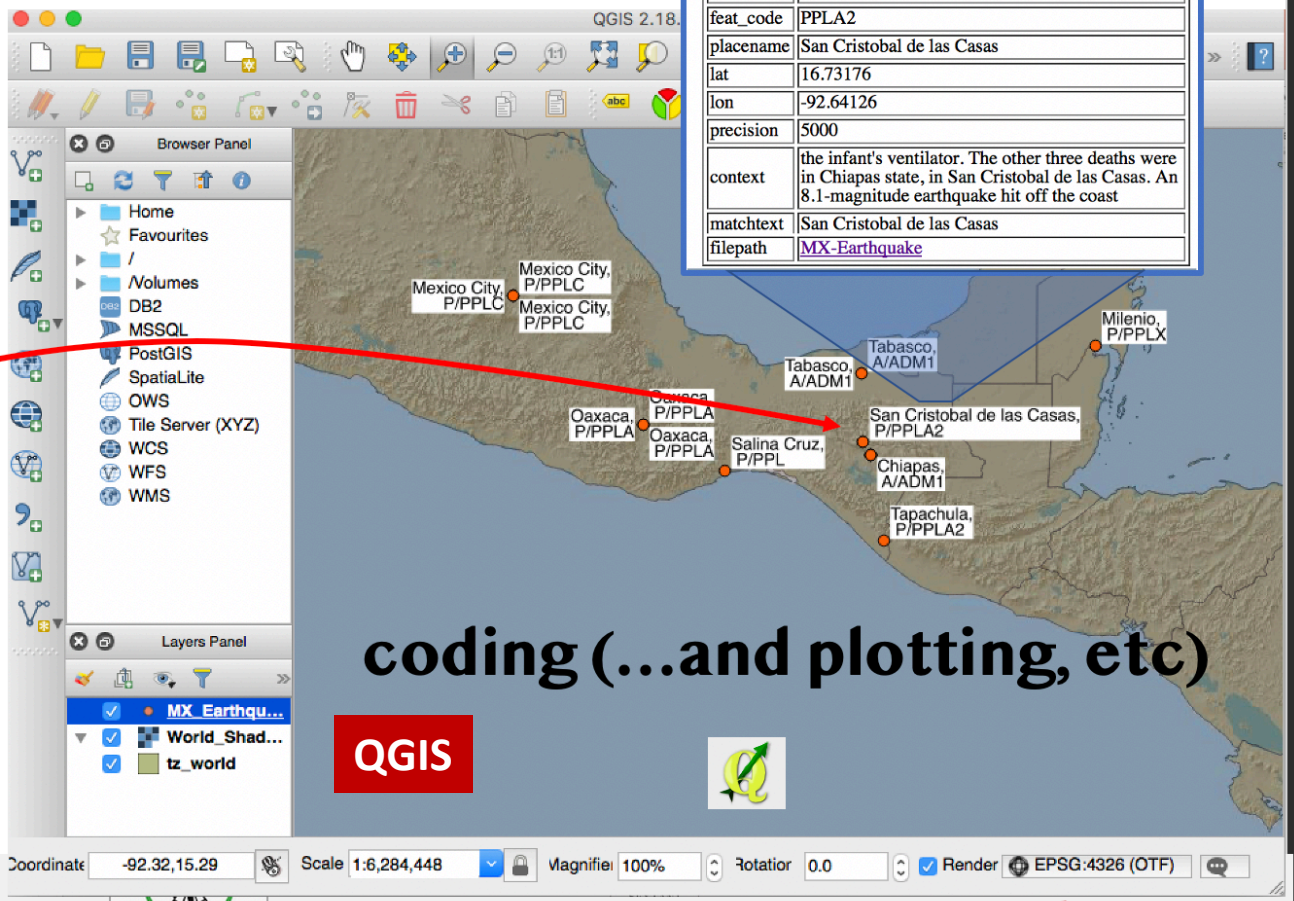
An 8.1-magnitude earthquake hit off the coast of southern Mexico, toppling houses in Chiapas state, causing buildings to sway violently as far away as the country's distant capital and setting off a tsunami warning.

....

MITRE Annotation Toolkit (MAT)

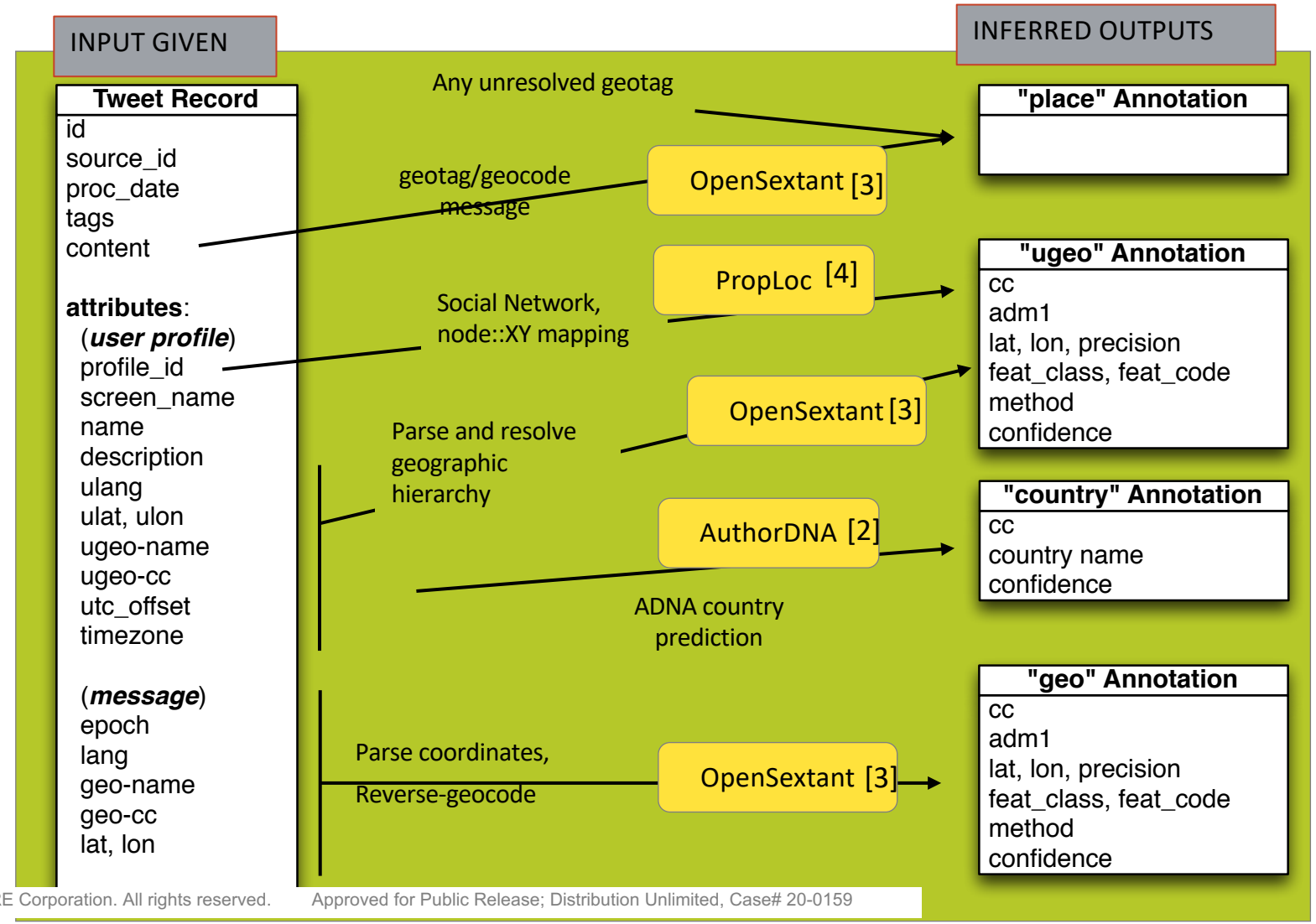
San Cristobal de las Casas

id	34
iso_cc	MX
province	05
feat_class	P
feat_code	PPLA2
placename	San Cristobal de las Casas
lat	16.73176
lon	-92.64126
precision	5000
context	the infant's ventilator. The other three deaths were in Chiapas state, in San Cristobal de las Casas. An 8.1-magnitude earthquake hit off the coast
matchtext	San Cristobal de las Casas
filepath	MX-Earthquake



Geo-inferencing Framework and Pipeline ***SocGeo***

SocGeo[1] = derive the best possible geographic characterization of social media data sets through many techniques applied to any part of the signal



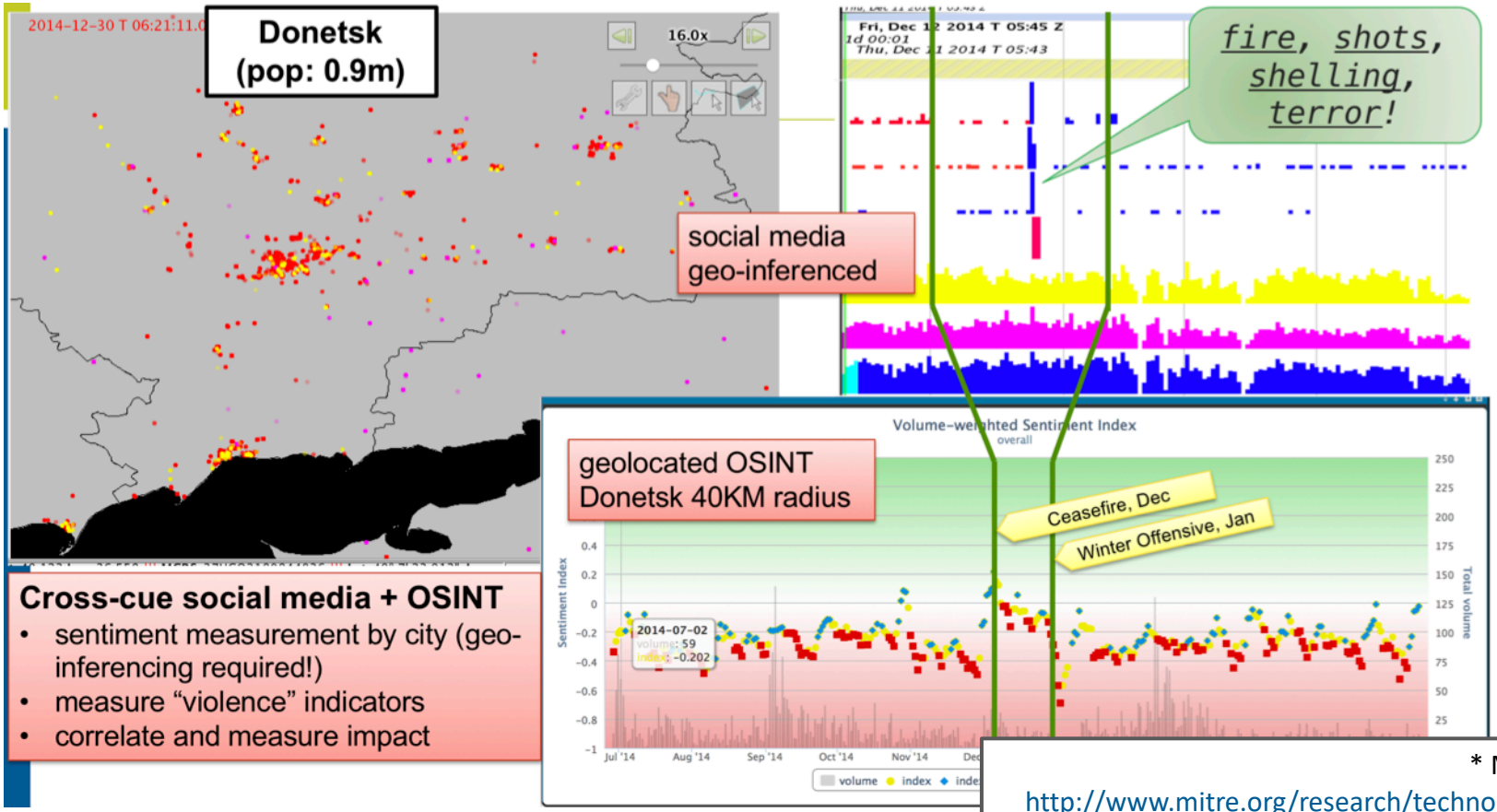
MITRE

SocGeo References

1. MITRE Social Radar, 2010-2017. SocGeo is a part of this larger portfolio of R&D. Technology sharing information: http://www.mitre.org/research/technology-transfer/technology-licensing/social-radar-technologies
2. MITRE, 2011. “Discriminating Gender on Twitter”, May 2011. https://www.mitre.org/publications/technical-papers/discriminating-gender-on-twitter
3. Ubaldino, 2019. OpenSextant “social geo” API, demo and example routines in “**org.opensextant.extractors.geo.social”** package https://opensextant.github.io/Xponents/doc/sdk-apidocs/.
4. Jurgens, 2013 . “That’s What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships”, HRL Laboratories. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6067/6366

Integration Time: A Commissioned Ukraine Study

Highlight the value of real-time analytics with open sources and social media as a result of spatio-temporal analytics augmented with native language processing. This analysis was intended to parallel study: Dr. Karber & Gen. Zamana. "*Hybrid Warfare in Ukraine Lessons Learned*": http://nssp.unm.edu/tegnelia-november-21-presentation.pdf



Cross-cue social media + OSINT

- sentiment measurement by city (geo-inferencing required!)
- measure "violence" indicators
- correlate and measure impact

MITRE Prototypes*

ISR Forensics

Sentimedir

SocGeo

OpenSextant

SascaNLP

DeepEye / Sentix

MITRE ATS Lab (Evaluation resources)

- SYSTRAN
- MongDB
- Twitter / Gnip
- Open Source News

* More info on select prototypes: http://www.mitre.org/research/technology-transfer/about-fastlicense

MITRE

Conclusion of OpenSextant

Extracting credible geographic information from unstructured data & text is hard!

- **Linguistic, cultural, historical reference information** is necessary, but a combination of these resources + advancements in NLP and AI may offer better, reasonable solutions. E.g., to get past limitations of either approach alone.
- ***Reasonable*** solutions is relative to end-user: Mobile app user vs. scientist in the field; developer vs. computational linguist, big data solution vs. desktop app, etc.
- Unstructured data processing is guided by many standards, yet **geo-inferencing of such data** remains an art, with much less of a clear expectation of behaviors and outputs from geotaggers, geocoders, etc. **OpenSextant** offers a variety of best practices in terms of Gazetteer curation, text processing, evidence-based disambiguation, and use of standards in geocoding
  - **Recommendation**: Convene a focus group of users, geographers, developers and linguists at GIS, NLP or AI conference or formalize an open source community
- Acceleration in this field comes from improved **open source collaboration**, **sharing of research**, and **publishing of software artifacts**.

OpenSextant is open source that gets to the heart of your geospatial discoveries in free text.

This is a quest for truly "Geospatially Literate" software

 Approved for Public Release; Distribution Unlimited, Case# 20-0159

MITRE

Possible Future Collaborations

- Participate in Out of Eden Walk (https://gis.harvard.edu/out-eden-walk)
  - Support “news triage” geospatially
  - Render existing OOEW articles and posts into usable map
  - Discover related places or unknown places i.e., known place not in official gazetteer working in a language other than English
  - Demonstrate “open source” high-tech geography concepts to students, teachers and others outside of the geographer/data science realm
- Research Collaboration on Challenges above or related ones.
- Use OpenSextant: See Docker page https://hub.docker.com/r/mubaldino/ for using RESTful web-service
- Contribute to or Develop with OpenSextant: https://opensextant.github.io/Xponents/

Contact: Marc Ubaldino
ubaldino@mitre.org

MITRE

The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our federally funded R&D centers and public-private partnerships, we work across government to tackle challenges to the safety, stability, and well-being of our nation.

Learn more www.mitre.org

 Approved for Public Release; Distribution Unlimited, Case# 20-0159