LUCENE / SOLR REVOLUTION / 2017

SEPTEMBER 12-15, 2017
LAS VEGAS, NV

Discovering World Geography in Your Data

Marc Ubaldino
Co-founder, OpenSextant.org
MITRE Corporation

Agenda

- OpenSextant – a family of extraction tools, using Solr
- Drivers for geotagging, etc.
- Advantages of Solr and SolrTextTagger
- In action
- Applications for you

OpenSextant: Geotagging and more

One project to consolidate our best solutions and techniques for geotagging

http://opensextant.org

MITRE open source since 2013

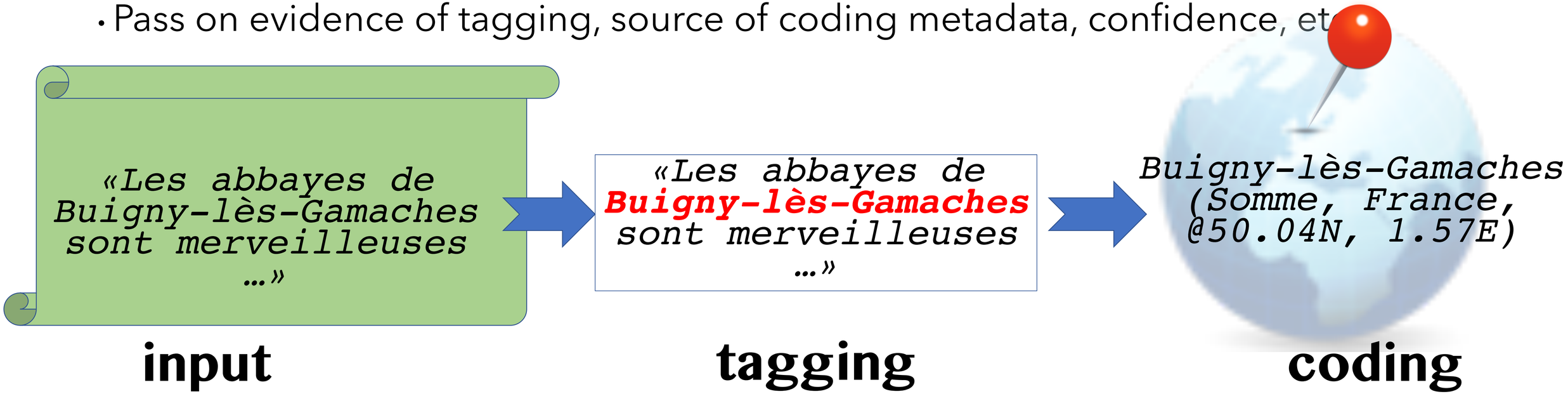
… still active today

Related Outcomes

- Gazetteer curation
- SolrTextTagger
- Related extractors: date/time, coordinates, patterns, vocab
- Formal uses of Tika
- Simplify GIS uses

Geotagging & Geocoding Primer

- Acquire text, detect language
- Identify named or patterned location mentions
- Assign location and metadata to mentions
- Pass on evidence of tagging, source of coding metadata, confidence, etc



Open SEXTANT

Drivers for Geotagging

...and other language technologies

- **Volume of reference data** or model size is large
- **Speed**
- **Localization** (language, colloquialisms, culture)
- Best **quality** & most **current** data may be online solution in production
- **Cost** of on-site solutions, cost of R&D

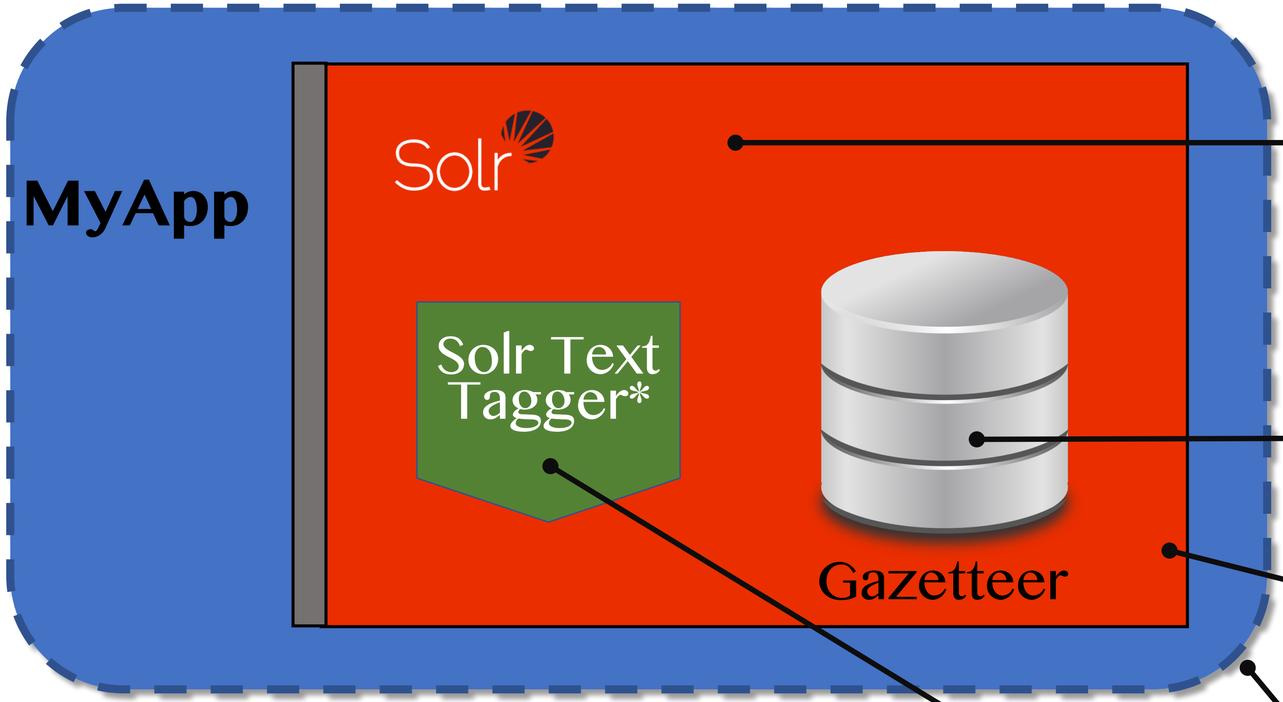
17 million place names for 9.5 million locations

Half of the world speaks 6 languages.

Connectivity is assumed

Advantages of OpenSextant using SolrTextTagger

Operates on laptop, server or cloud



Capability of Lucene, Power of Solr:
- Tokenizers & resources for dozens of languages
- Finite State Transducer (FST)

Compact! 17 million named places:
2.5 GB index on disk
~1.0 GB in RAM

Fast! Tag ~50 KB/sec

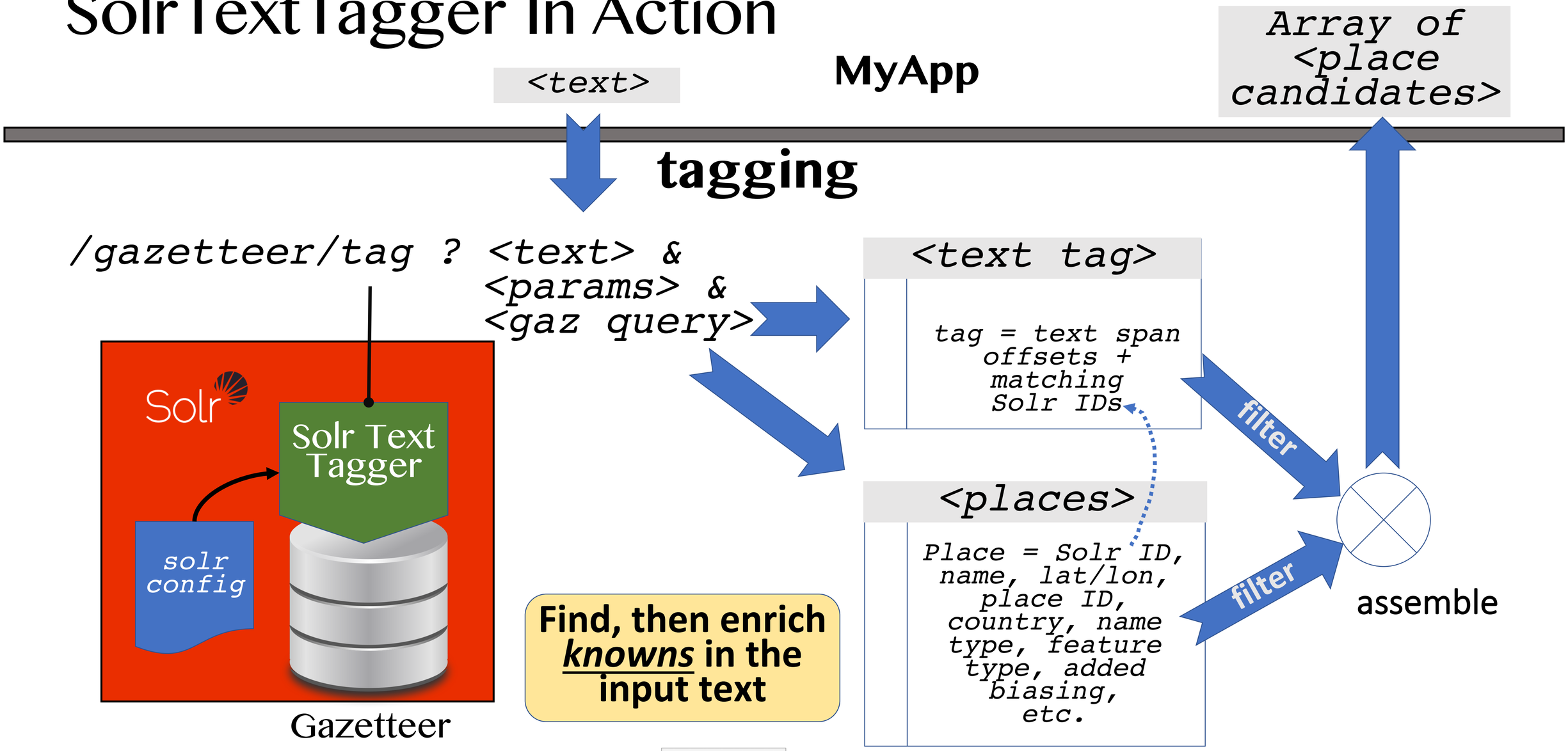
No network connection required at runtime with EmbeddedSolr

Easily built using Maven

Flexible and extensible

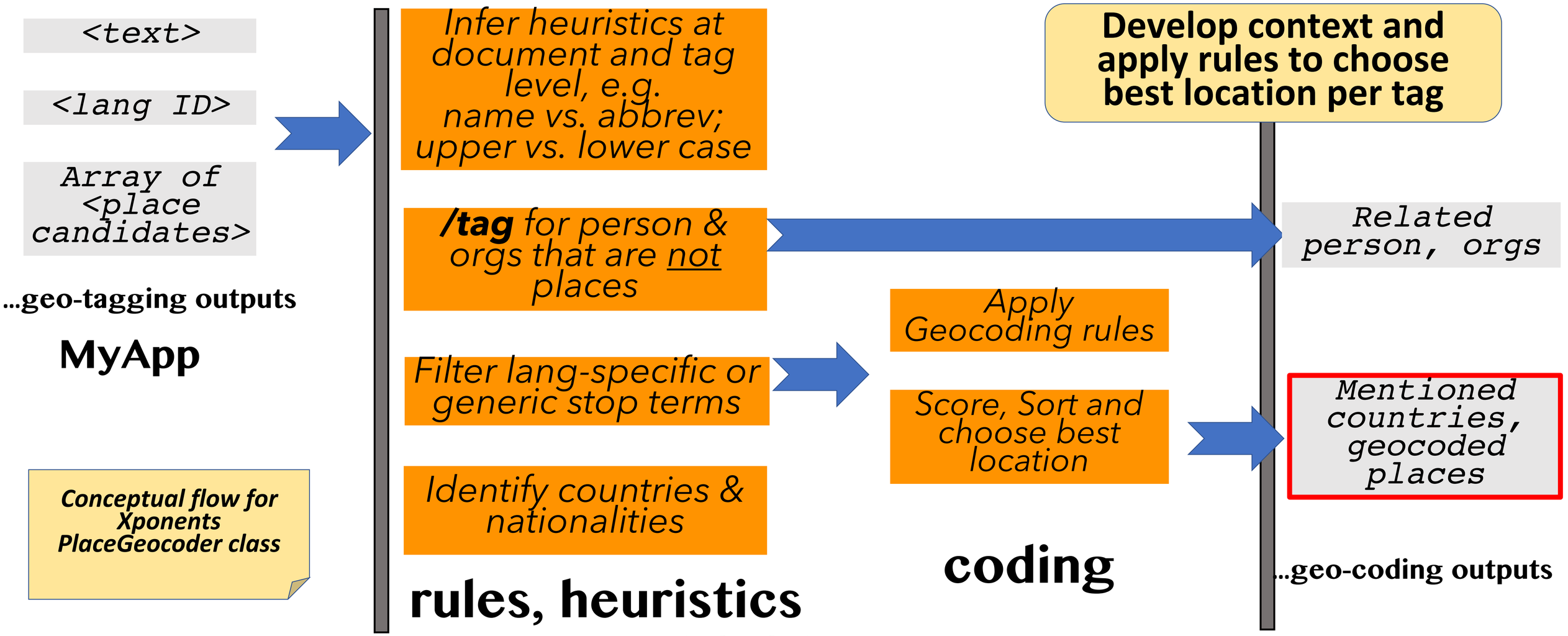
Well-documented

SolrTextTagger In Action



Open SEXTANT

Using output from SolrTextTagger



Solr, Lucene & SolrTextTagger Roles in Tagging

These all maximize & refine tagger recall, but can yield excessive noise.

Geocoding steps balance noise using filters, rules and heuristics to improve precision

- Solr provides
  - **/tag** Request Handler configuration
  - **/select** Request Handler, which is applied here to geo-name search, spatial proximity search, other parametric searches
  - Configurability
  - **Scripting**, e.g., gazetteer index creation involves heavy scripting
- Lucene provides
  - **Tokenization analyzer**: standard, Arabic, Asian languages
  - **Filters**: char mapping diacritics, lower case, etc.
- SolrTextTagger provides
  - Mapping Solr **indexed schema field** to FST build at runtime
  - **Filter query** to pull the right data into FST

Components to Build Applications

- Project: https://github.com/OpenSextant
  - **Xponents** – extractors
  - **OpenSextantToolbox** – GATE-based extractors
  - **Gazetteer** – gazetteer data curation
  - **SolrTextTagger** – flexible, FST tagger
  - GISCore – spatial data formatting

- Modes of use in **Xponents**:
  - Java API, e.g., EmbeddedSolr
  - REST via Restlet ™, a.k.a., Xponents Xlayer
  - Solr Admin
  - Map Reduce

Application Domains for You

- **Tourism**: Find places favored by tourists or businesses
- **Park Service**: Discover hot spots of interest in your parks
- **Humanitarian**: Track remote villages in need
- **Education**: Foster geographic and cultural literacy
- **Security**: Monitor serious regional natural or human events
- **Discovery**: Mine noisy public data for geographic cues in unfamiliar languages or cultures
- More!

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.

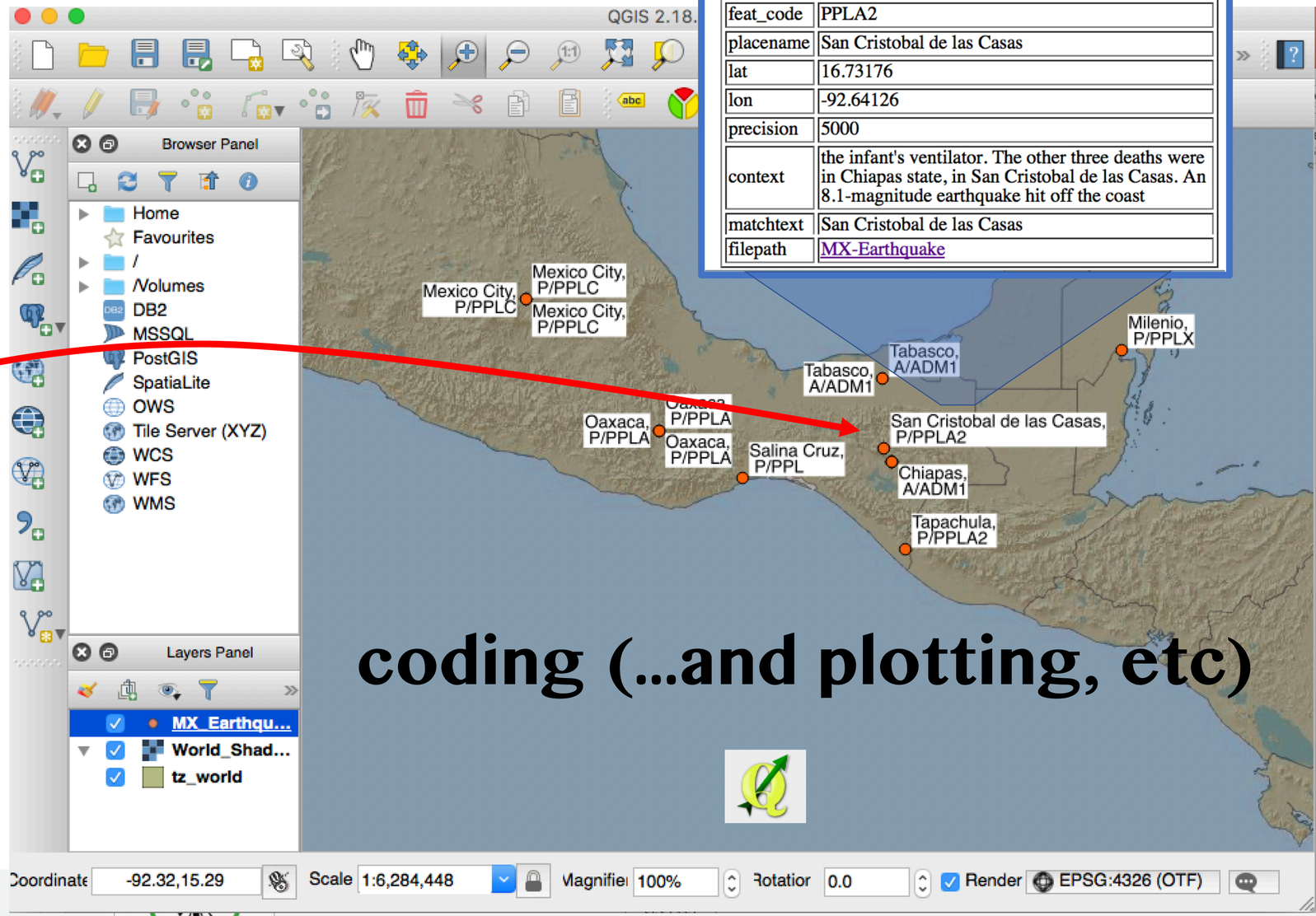
An 8.1-magnitude earthquake hit off the coast of southern Mexico, toppling houses in Chiapas state, causing buildings to sway violently as far away as the country's distant capital and setting off a tsunami warning.

....

tagging

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



coding (...and plotting, etc)

Summary

- **OpenSextant** benefits from Solr and **SolrTextTagger** because they are
  - **Flexible, compact and work well off-line**
  - **Minimize memory**
  - **Fast**
  - **Multi-lingual** support is superior
  - Reasonable balance between **cost, quality, performance**

OpenSextant helps you discover and resolve world wide geography in any text!

LUCENE / SOLR REVOLUTION / 2017

SEPTEMBER 12-15, 2017
LAS VEGAS, NV

Thank You

contact: ubaldino@mitre.org

Open SEXTANT

MITRE

Approved for Public Release
Distribution Unlimited
MITRE Case # 17-3145