MetaCarta GeoSearch Toolkit for Solr

James Goodwin
Principal Engineer, Nokia

Overview

- Introduction to MetaCarta
- About Nokia
- MetaCarta Geographic Search
- Defining GeoSearch Functionality for Solr
- Toolkit Flow
- The Components of the GeoSearch Toolkit for Solr
- Geographic Metadata Dynamic Fields
- Query Syntax
- Geographic Relevance
- Geographic Filter
- An Example Configuration
- A Brief Demonstration
- Current Status
- Questions / References

Introduction to MetaCarta

- A leader in geographic search technology since 2001
- Customers across private and public sectors:
  - Public: DIA Counter Terrorism, ICES, SOCOM, Australian Government
  - Civilian: USDA and Smithsonian
  - State & Local: North Central Texas Fusion Center
  - Energy: Schlumberger, Worldwide Reseller,Chevron, Shell, BP, Total, BHP Billiton, Anadarko
  - Media: BBC, Star-Telegram, Reuters, Society of Petroleum Engineers,Websites
  - Web Applications: YourStreet, PIIM
- Patented Geographic Search Technology
- Recently acquired as a wholly owned subsidiary by Nokia

About Nokia

We are hiring!

About Nokia

1.2 billion people are using Nokia devices worldwide

Devices sold in 220 countries/territories

Symbian available in 180 languages

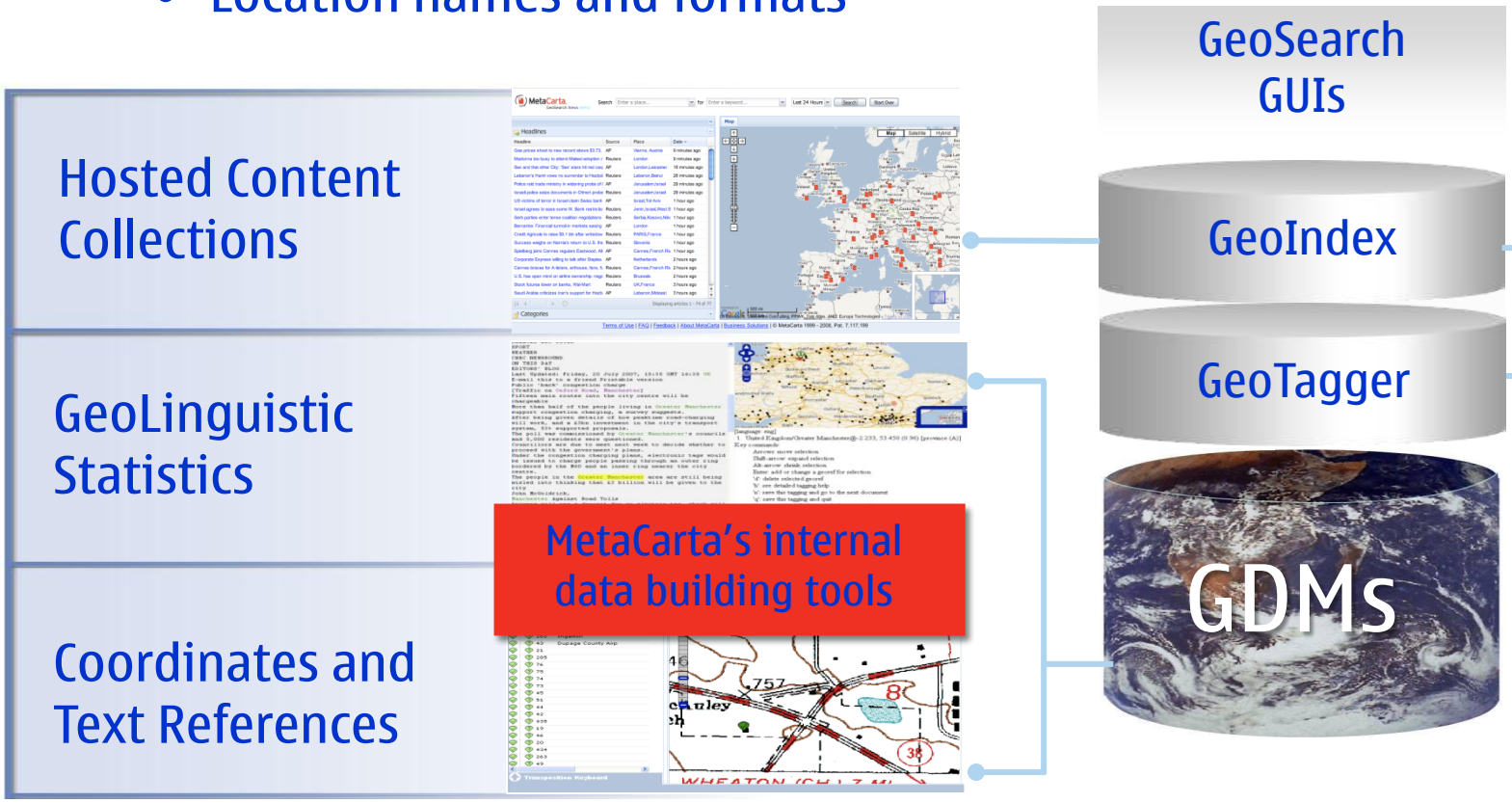
Since the start of this talk:

- 1K+ Nokia devices were made and sold (13/sec)
- 15M+ phone calls were made using Nokia phones
- 3M+ text messages were sent using Nokia phones

MetaCarta Geographic Search

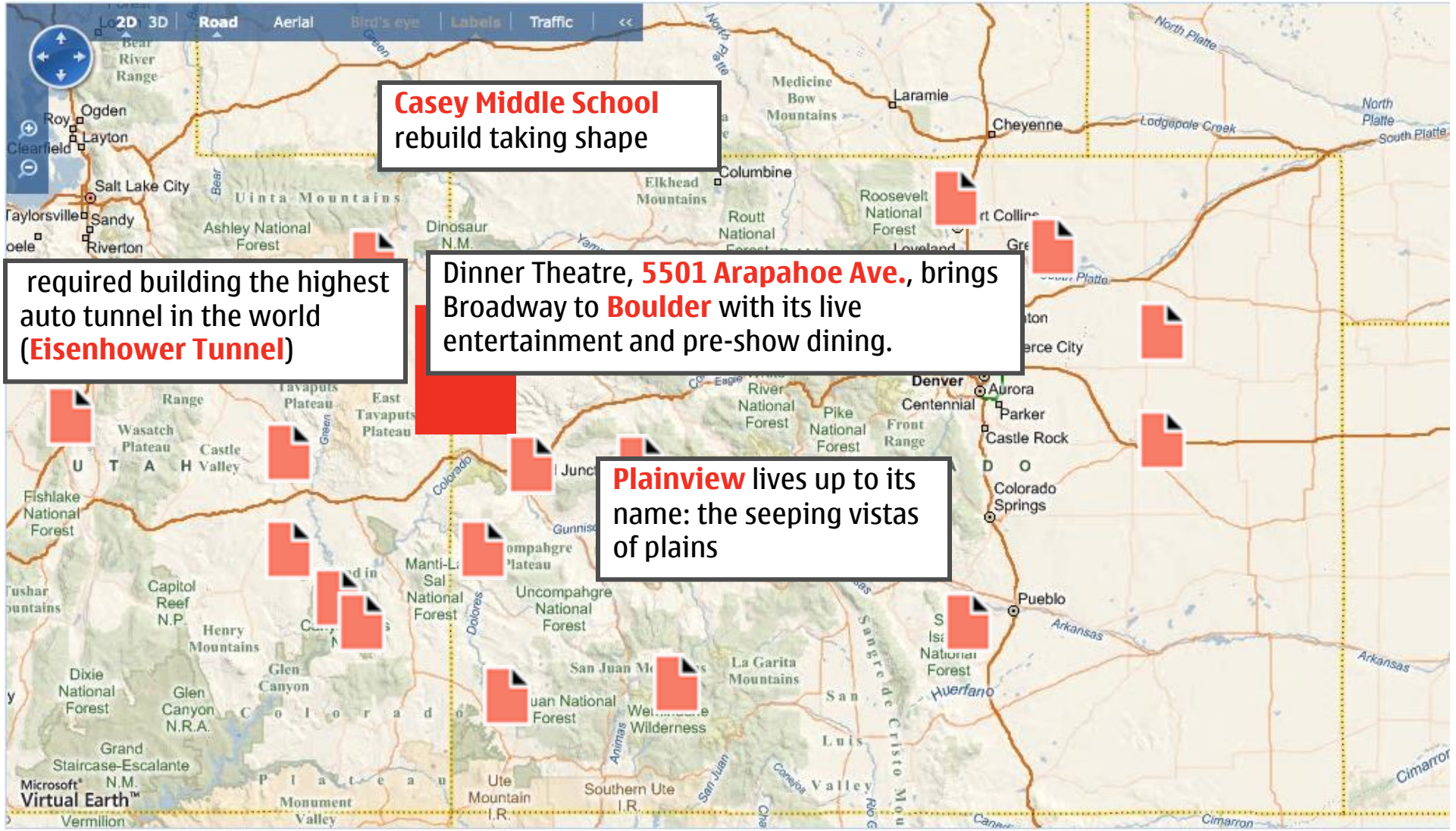
MetaCarta gathers and creates several forms of geo-textual data:

- Collections of GeoRelevant unstructured text
- Linguistic statistics from manually proofed texts in many genres and languages.
- Location names and formats



NOKIA

MetaCarta Geographic Search

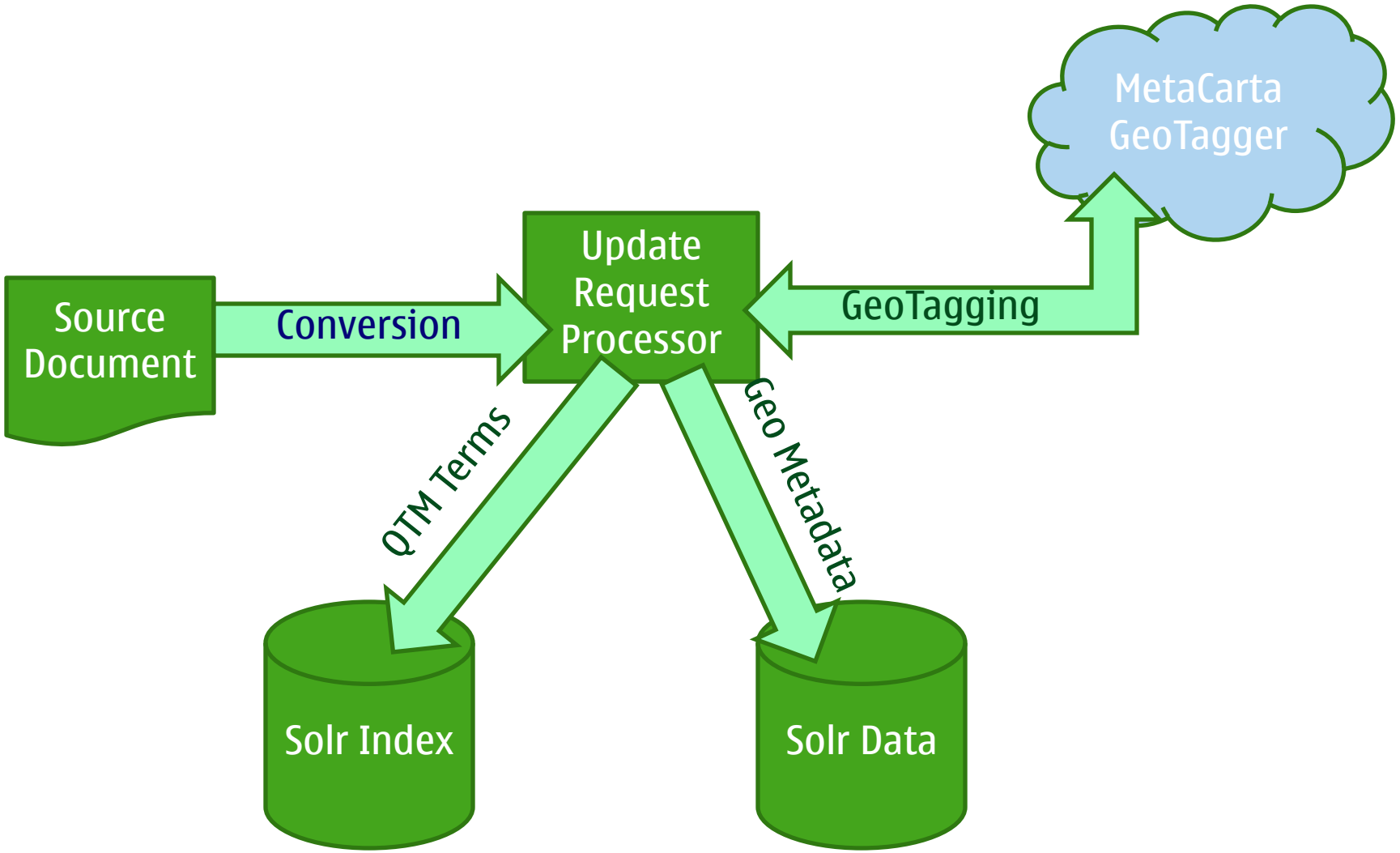


NOKIA

Defining GeoSearch Functionality for Solr

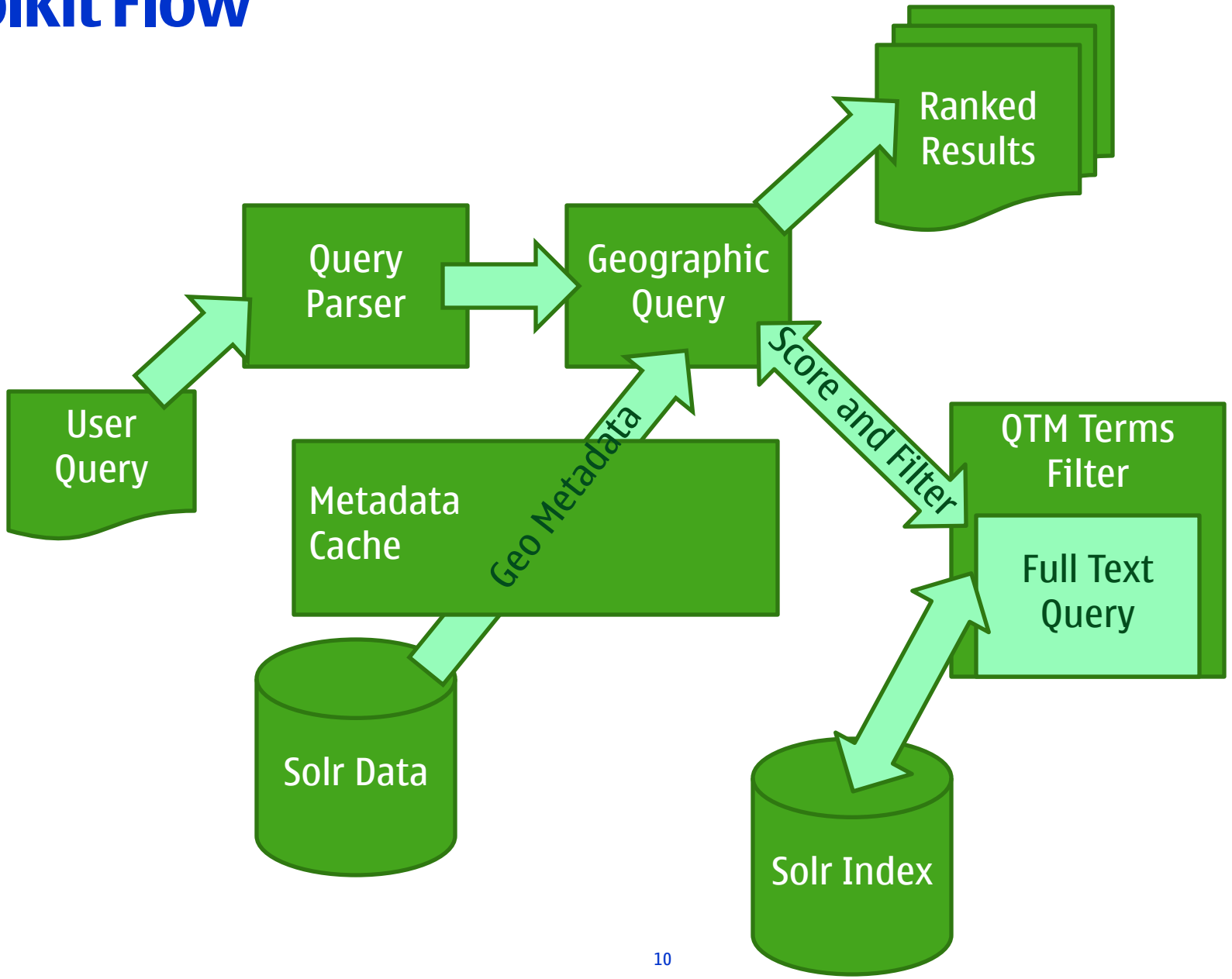
- Integrate with MetaCarta's GeoTagger service to tag geographic references in documents as they are being indexed
- Make tagged geographic references searchable at the same time as full text queries
- Express queries as a combined geographic and full text query
- Filter results based on a bounding box, confidence, and other geographic metadata
- Rank results based on geographic relevance

Toolkit Flow



NOKIA

Toolkit Flow



NOKIA

The Components of the GeoSearch Toolkit for Solr

- MetaCartaUpdateRequestProcessor
  - Plugs into Solr's update request processor chain
  - GeoTags configured fields in the document using MetaCarta GeoTagger (or reads GeoMarkup XML from a field)
  - Inserts dynamic fields into the document containing geographic metadata
  - Indexes a QTM or QuadTree code for each GeoTag
- MetaCartaQParserPlugin
  - Plugs in as a query parser
  - Supports geographic parameters
  - Delegates to any other query parser to construct the full text query
  - Generates a MetaCartaGeoQuery which implements filtering and ranking
- MetaCartaGeoQuery
  - Wraps another query
  - Boosts results based on geographic relevance
  - Filters results based on bounding box and confidence

Geographic Metadata Dynamic Fields

These are the sufixes of dynamic fields that are created based on each source document field selected for geographic data:

- **_mc_qtm** space separated qtm strings
- **_mc_latlon** pairs of latitude, longitude
- **_mc_pos** start of each geotag in the document
- **_mc_len** length of each geotag in the document
- **_mc_conf** confidence of each geotag in the document
- **_mc_wgt** weight of each geotag in the document

NOTE: the fields below this line are created optionally and are not required

- **_mc_name** the canonical name of the geotag
- **_mc_type** the type of the geotag
- **_mc_class** the class of the geotag
- **_mc_ctry** (country code,country name) pairs
- **_mc_prov** (province_code,province_name) pairs
- **_mc_cconf** country confidence
- **_mc_pconf** province_confidence
- **_mc_pop** population

Query Syntax

```
{!metacarta minlat=44.0 minlon=54.0
maxlat=66.0 maxlon=70.0 georel=true
minpop=50000} attr_content:school board
```

- Selects our query parser
- Bounding box parameters are specified
- Other options like using geographic relevance or filtering on population
- Followed by the syntax for the wrapped query parser
- Here it is the default Solr query parser

Geographic Relevance

- Boost added to full text relevance
- Terms closer to geographic reference get more boost
- Is a function of confidence of the geographic reference
- Is diluted when the distance to geographic reference is large
- Is diluted when the number of geographic references is high
- Only geographic references that are in the bounding box contribute to geographic relevance

Geographic Filter

- Bounding box is transformed into a minimal set of QTM prefixes
- A filter query on these prefixes against the QTM dynamic field matching the full text query field
- Initially retrieves only documents matching these codes
- MetaCartaGeoQuery's scorer acts as a detailed filter against the bounding box as well as against other geographic metadata
- Documents that have no geographic references with high enough confidence in the bounding box are discarded
- If extended geographic metadata is enabled, filtering on type, class, country, province metadata can also be used to filter

An Example Configuration

In solrconfig.xml configure the request processor chain and add it to a request processor:

```
<updateRequestProcessorChain name="metacarta">
    <processor
class="com.metacarta.GeoSearch.MetaCartaUpdateRequestProcessorFactory">
      <str name="geoTaggerHost">localhost</str>
      <str name="geoFields">attr_content</str>
      <bool name="geoMetadata">true</bool>
    </processor>
    <processor class="solr.LogUpdateProcessorFactory" />
    <processor class="solr.RunUpdateProcessorFactory" />
  </updateRequestProcessorChain>

<requestHandler name="/update/extract"
class="org.apache.solr.handler.extraction.ExtractingRequestHandler"
startup="lazy">
    <lst name="defaults">
        ...
      <str name="update.processor">metacarta</str>
    </lst>
  </requestHandler>
```

An Example Configuration

In solrconfig.xml configure the query parser and the cache:

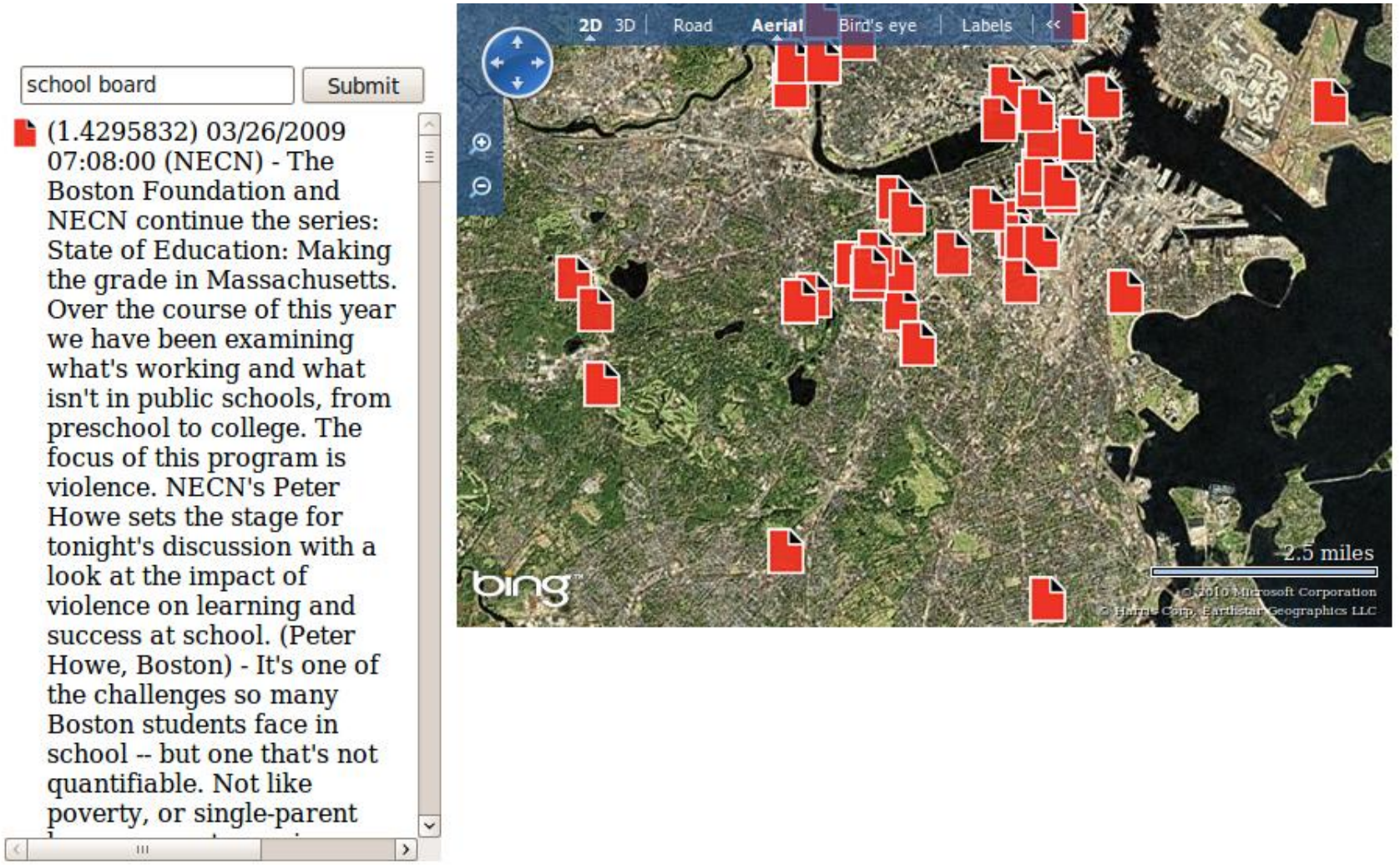
```
<cache name="mcCache" class="solr.LRUCache" size="4096"
initialSize="1024" />

<queryParser name="metacarta"
class="com.metacarta.GeoSearch.MetaCartaQParserPlugin">
      <str name="geoFields">attr_content</str>
      <str name="latlonField">attr_latlon</str>
</queryParser>
```

An Example Configuration

In schema.xml configure field types and dynamic fields:

```
<fieldtype name="mcdarray" class="com.metacarta.GeoSearch.MetaCartaDoubleArrayField" />
<fieldtype name="mciarray" class="com.metacarta.GeoSearch.MetaCartaIntArrayField" />
<dynamicField name="*_mc_qtm" type="textgen" indexed="true" stored="true" multiValued="true" />
<dynamicField name="*_mc_pos" type="mciarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_len" type="mciarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_latlon" type="mcdarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_conf" type="mcdarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_wgt" type="mcdarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_name" type="textgen" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_class" type="textgen" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_type" type="textgen" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_ctry" type="textgen" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_prov" type="textgen" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_cconf" type="mcdarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_pconf" type="mcdarray" indexed="false" stored="true" multiValued="true" />
<dynamicField name="*_mc_pop" type="mciarray" indexed="false" stored="true" multiValued="true" />
```

A Brief Demonstration



Current Status

- Private alpha release
- Refining performance and storage overhead
- Tuning ranking

Questions / References

E-mail: james.2.goodwin@nokia.com

Web: http://www.metacarta.com, http://ovi.nokia.com/services/

Thanks to: David Smiley at MITRE for his help testing and improving the toolkit. Check out his book: **Solr 1.4 Enterprise Search Server**

QTM reference: "Encoding and Handling Geospatial Data with Hierarchical Triangular Meshes" (http://www.spatial-effects.com/papers/conf/GDutton_SDH96.pdf)

QuadTree reference: "QUADTREE ALGORITHMS AND SPATIAL INDEXES" (http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u37.html#UNIT37)

NOKIA

NOKIA