Exploiting Open-Source Geospatial Data for the Maintenance of GIS Databases

Barry Bitters Ph.D., GISP
Leidos, Inc. (Formerly SAIC)

Images courtesy of Google Earth™

Location-Based Applications

- **Require accurate and up-to-date cultural and landmark feature location information**
  - **Traditional applications:**
    - **Cartographic**
    - **Intelligence**
  - **More recent applications:**
    - **Mobile location-based services**
    - **Real-time locating systems**

Open-Source Data

- **It's available on Internet – So it must be true??**
- **It's absolute truth when verified on the ground**
- **It's probably true when:**
  - **Verified on imagery**
  - **Referenced by multiple sources**
  - **By attribute comparison - Majority rule techniques**

Repurposing Web Data

- **Mashups and Extraction, Transformation, Loading (ETL) Software**

- **Reuse of existing data, many times propagates errors, omissions and commissions**

Typical Data Problems

- Feature Names
- Duplicate Records
- Omitted Data
- Faulty Location Data
- Attribute Inconsistencies
- Human Interaction Induced Errors
- Intentional Errors

The Problem…

- **The WWW contains a wealth of geospatial data**
  - **Produced by a variety of sources and methods**
  - **Available in varying degrees of:**
    - **Currency**
    - **Reliability**
    - **Accuracy**
    - **Precision**
- **How to rapidly verify & validate the existence of cultural features referenced on the Web?**
- **Is it economically feasible???**

Basic Assumptions

- **Feature database maintenance – costly; a forever endeavor**
- **Feature databases are never 100% complete**
- **Feature databases are never 100% accurate**
- **Augmenting existing databases with open-source data can:**
  - **Improve currency.**
  - **Improve depth of detail.**
- **Google Earth™ (the commercial version) can serve as a worldwide geographic coordinate reference**

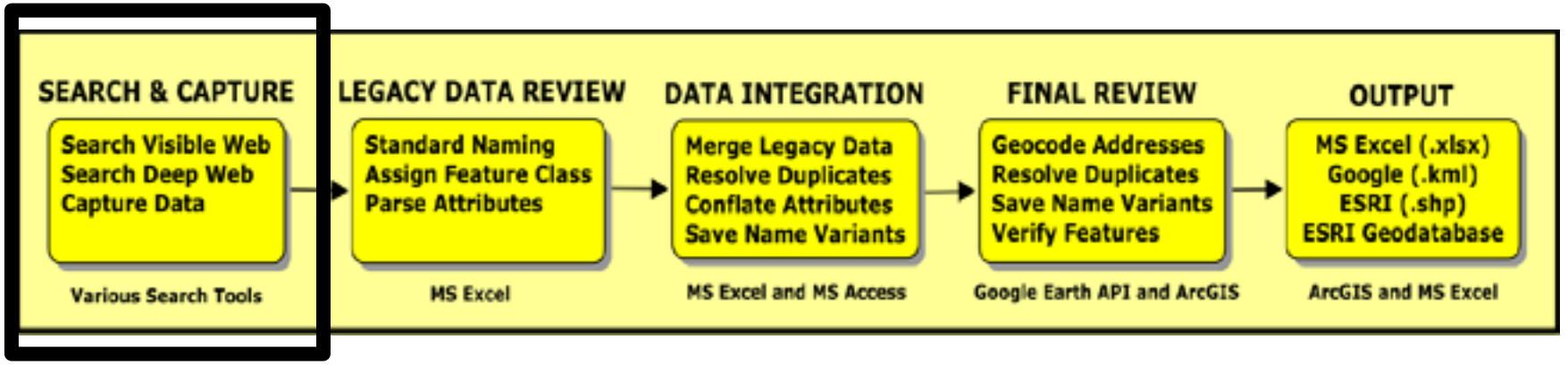
Area of Interest

- **Why South Africa??**
  - **Never visited - therefore no bias**
  - **Emerging nation status**
- **Cultural Features of Interest**
  - **Airfields**
  - **Golf courses**
  - **Police stations**
  - **Post offices**
  - **Maritime lights**
  - **In-Work – schools, churches, cemeteries, courthouses, towers, tanks**

Processing Priorities

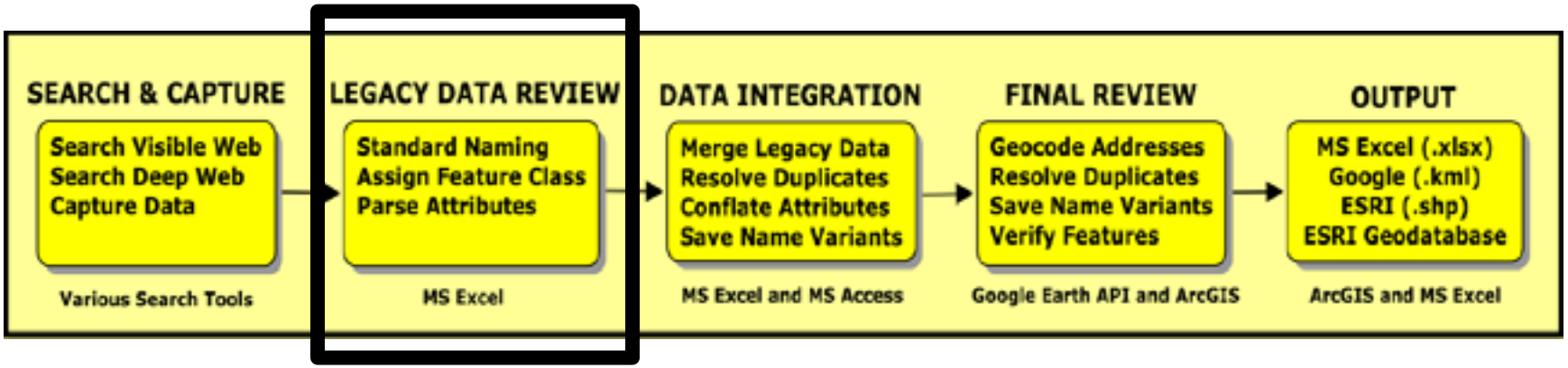
- Capture ALL pertinent open-source data
- Identify/create a “standard nomenclature”
- Merge disparate open-source records
- Resolve attributes
- Cull duplicate records
- Preserve name variants
- Maintain record-level legacy metadata
- Verify feature existence
- Refine geo-positions
- Output in a variety of formats

Search & Capture Open-Source Data



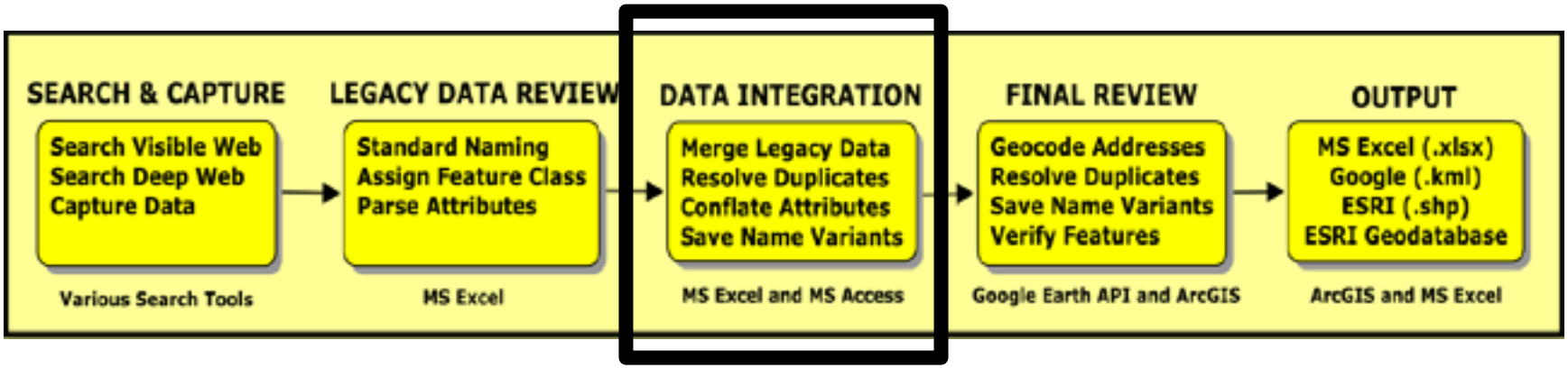
- Identify Relevant Geospatial Data
  - Visible Web
  - Deep Web (aka. Hidden Web)
- Categories of open-source data
  - Structured data – Formatted, e.g. spreadsheets, databases
  - Semi-structured data – Semi-formatted, e.g. Web tables
  - Unstructured data – Free text
- Each category requires different data capture approaches

Legacy Data Review



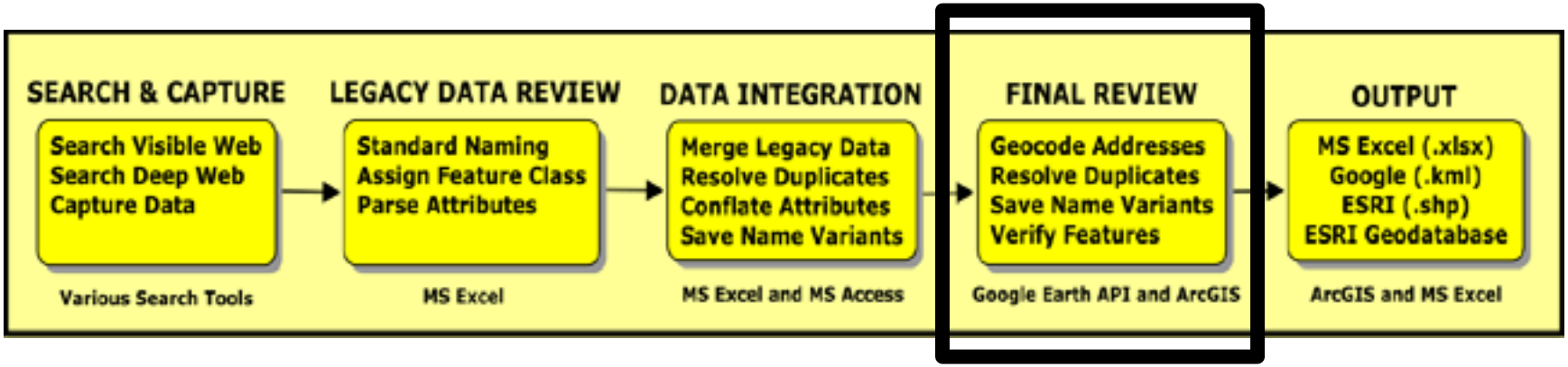
- Normalize attributes
- “Standardize” feature names
- Assign feature classes
- Record level metadata – legacy source & DOI
- Parse attributes
- Convert geo-coordinates to decimal degrees WGS-84

Legacy Data Integration



- **Merge all legacy data records**
- **Conflate attributes**
- **Resolve duplicates – first phase**
- **Preserve name variants**

Final Review



- Geocode addresses
- Final culling of duplicates
- View features on imagery (ground and/or vertical)
  - Verify existence
  - Verify function
  - Validate name (when possible)
- Preserve name variants

Results

- **Over 5,000 feature instances reviewed on ground and/or vertical imagery**
- **90% now have refined geo-positions based on commercial Google Earth™ imagery**
- **Added Value**
  - § “Opportunity Collects” – additional feature instances captured during final image review
  - § Detailed lists of image signatures - Regional/Country Specific & Feature Class Specific

Legacy Data Merging Results

Feature Type	*No. of Legacy Sources*	*Largest Legacy Source*	*Final Dataset Size*	*% Increase*	*** New Features*
Airfields	**17**	**592** (NGA GNS)	**1102**	**46**	**259**
Golf Courses	**8**	**461** (Golf World)	**577**	**20**	**14**
Police Stations	**8**	**1146** (SAPS)	**1214**	**6**	**6**
Post Offices	**7**	**2444** (SA Travel)	**2787**	**12**	**0**
Maritime Lights	**8**	**172** (NGA Lights)	**229**	**25**	**12**

NGA GNS - NGA Geonames Server - http://earth-info.nga.mil/
Golf World - Golf World Map - http://www.golfworldmap.com/
SAPS - South African Police Service - http://www.saps.gov.za/
SA Travel - South African Travel Directory - http://sa.travel-directory.co/
NGA Lights - NGA List of Lights - http://msi.nga.mil/

** **Opportunity Collects - Features found on imagery but not referenced in any legacy sources.**

Feature Verification Results

Feature Type	*Final Dataset Size*	*Percent Image Verified*	*Percent Probable*	*Percent Possible*	*Percent Unverified*
Airfields	**1102**	**78.5**	**0.0**	**2.9**	**18.6**
Golf Courses	**577**	**97.9**	**0.0**	**0.6**	**1.5**
Police Stations	**1214**	**55.2**	**20.7**	**14.4**	**9.7**
Post Offices	**2587**	**50.0**	**15.4**	**23**	**11.6**
Maritime Lights	**229**	**90.9**	**0.0**	**3.0**	**5.7**

In Closing...





Barry Bitters
6731 Avenida de Galvez
Navarre, Florida 32566
Ph. [850] 684-3052
bittersb@gmail.com
bittersb@leidos.com

Images courtesy of Google Earth™

Data Processing Priorities

- Capture ALL pertinent open-source data
- Initially, preserve ALL feature instances.
- Initially, preserve ALL attributes of each feature instance
- Identify/create a “standard nomenclature”
- Define field names
- Conflate disparate open-source records
- Merge attributes
- Cull duplicate records but preserve ALL attributes
- Maintain record-level legacy metadata
- Insure all feature instances have “precise” and “accurate” geo-coordinates
- Verify feature existence
- Output in a variety of formats

Security Considerations

- **Data capture and image verification demand an Unattributed & Unfirewalled workstation!**
  - **The commercial version of Google Earth is an open line visible to the world.** **"**Everyone" potentially can see geocoding operations, place names and zooming to point locations
  - **Web scrappers and search engines generate a visible trail on the internet which "everyone" can potentially see**
    - The Onion Router (TOR) Browser – is a secure Web browser that provides encrypted anonymity while interacting on the Web
    - FBI and NSA will take a keen interest in TOR Browser users

Evidence-Based Verification

- **Evidence used to evaluate and verify an object's existence (ordered from most to least reliable):**
  - Ground-based imagery verification
  - Vertical imagery verification
  - Identification of key visual signatures viewed on ground-based imagery
  - Identification of key visual signatures viewed on vertical imagery
  - Referenced in multiple authoritative sources
  - Referenced in multiple credible sources
  - Referenced in a single authoritative source
  - Referenced in a single credible sources
  - Referenced in a single non-authoritative source
  - Purely image analysts "gut instinct"

Who Am I?

- **20 years – US Army – Image Intelligence, Military Topography and Terrain Analysis.**
- **13 years – Lockheed (AFSOC) – Flight simulation database generation.**
- **5 years – University of West Florida – Environmental and GIS instruction & geospatial research.**
- **7 years – SAIC/Leidos (DIA, NGA & IARPA) – ArcGIS/Oracle database development, object-based feature extraction, geospatial semantics, 3D-stereo visualization, geospatial instruction.**