Reconciliation with Elasticsearch

Nathan Lomeli
What is Elasticsearch?

Open Source, Distributed, RESTful Search Engine  
https://www.elastic.co/products/elast...
Rising popularity
Data in, data out

PUT /megacorp/employee/1
{
   "first_name" : "John",
   "last_name" : "Smith",
   "age" : 25,
   "about" : "I love to go rock climbing",
   "interests" : [ "sports", "music" ]
}
Data in, data out

PUT /megacorp/employee/1
{
    "first_name" : "John",
    "last_name" : "Smith",
    "age" : 25,
    "about" : "I love to go rock climbing",
    "interests" : [ "sports", "music" ]
}

PUT /megacorp/employee/2
{
    "first_name" : "Jane",
    "last_name" : "Smith",
    "age" : 32,
    "about" : "I like to collect rock albums",
    "interests" : [ "music" ]
}
GET /megacorp/employee/_search?q=last_name:Smith
{
    "hits": {
        "total": 2,
        "max_score": 0.30685282,
        "hits": [
            {
                "_source": {
                    "first_name": "John",
                    "last_name": "Smith",
                    "age": 25,
                    "about": "I love to go rock climbing",
                    "interests": ["sports", "music"]
                }
            },
            {
                "_source": {
                    "first_name": "Jane",
                    "last_name": "Smith",
                    "age": 32,
                    "about": "I like to collect rock albums",
                    "interests": ["music"]
                }
            }
        ]
    }
}
Multi-field search

GET /megacorp/employee/_search

{
    "query" : {
        "bool" : {
            "must" : {
                "match" : {
                    "last_name" : "smith"
                }
            },
            "filter" : {
                "range" : {
                    "age" : { "gt" : 30 }
                }
            }
        }
    }
}
When can Elasticsearch help with reconciliation?

- good when source data is labeled or structured
- good when multiple fields aid discrimination
- good when target vocabulary has relevant metadata fields
- good for providing ranked list of candidate matches
- good for labor saving compared to manual reconciliation
Application: The Provenance Project
**Goal: structure written records**
The semantic model
Purpose: to explore provenance by entity
Means: digitizing records ...

<table>
<thead>
<tr>
<th>artist_authority_artist_1</th>
<th>variant_names_artist_1</th>
<th>birth_date_artist_1</th>
<th>death_date_artist_1</th>
<th>period_activ_century_activ</th>
<th>active_city_date_artist_1</th>
<th>nationality_artist_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLORI, ALESSANDRO</td>
<td>Allori, Alessandro; Alessandro Allori;</td>
<td>1535</td>
<td>1607</td>
<td>x</td>
<td>16th</td>
<td>Italian</td>
</tr>
<tr>
<td>ALLORI, ALESSANDRO</td>
<td>Allori, Alessandro; Alessandro Allori;</td>
<td>1535</td>
<td>1607</td>
<td>x</td>
<td>16th</td>
<td>Italian</td>
</tr>
<tr>
<td>MAILLOL, ARISTIDE</td>
<td>Maillo, Ariste; Ariste Maillot; Maillo; Maillot; Ariste Maillo; Maillot; Maillot</td>
<td>1861</td>
<td>1944</td>
<td>x</td>
<td>19th-20th</td>
<td>French</td>
</tr>
<tr>
<td>COROT, JEAN BAPTISTE CAMILLE</td>
<td>Corot, Jean Baptiste Camille; Jean-ba</td>
<td>1796</td>
<td>1875</td>
<td>x</td>
<td>19th</td>
<td>French</td>
</tr>
<tr>
<td>CHELMINSKI, JAN</td>
<td>Chelminski, Jan; chelminski j v; cheln</td>
<td>1851</td>
<td>1925</td>
<td>x</td>
<td>19th</td>
<td>Polish</td>
</tr>
<tr>
<td>PEALE, CHARLES WILLSON</td>
<td>Peale, Charles Willson; Peale; peale c</td>
<td>1741</td>
<td>1827</td>
<td>x</td>
<td>18th</td>
<td>American</td>
</tr>
<tr>
<td>PEALE, CHARLES WILLSON</td>
<td>Peale, Charles Willson; Peale; peale c</td>
<td>1741</td>
<td>1827</td>
<td>x</td>
<td>18th</td>
<td>American</td>
</tr>
<tr>
<td>BENBRIDGE, HENRY</td>
<td>Benbridge, Henry; Henry Benbridge;</td>
<td>1743</td>
<td>1812</td>
<td>x</td>
<td>18th</td>
<td>American</td>
</tr>
<tr>
<td>SULLY, THOMAS</td>
<td>Sully, Thomas; Thomas Sully; Sully; s</td>
<td>1783</td>
<td>1872</td>
<td>x</td>
<td>19th</td>
<td>American</td>
</tr>
<tr>
<td>CHASE, WILLIAM MERRITT</td>
<td>Chase, William Merritt; William Merr</td>
<td>1849</td>
<td>1916</td>
<td>x</td>
<td>19th</td>
<td>American</td>
</tr>
<tr>
<td>REMBRANDT, HARMENSZ. VAN RIJN</td>
<td>Rembrandt; Harmensz. van Rijn; Rem</td>
<td>1606</td>
<td>1669</td>
<td>x</td>
<td>17th</td>
<td>American</td>
</tr>
<tr>
<td>BERTZIG, ANTON</td>
<td>Bertzig, Anton; Anton Bertzik; bertsic</td>
<td>1856</td>
<td>1912</td>
<td>x</td>
<td>19th-20th</td>
<td>Austrian</td>
</tr>
<tr>
<td>TOMSON, CLIFTON</td>
<td>Tomson, Clifton; clifton tomsor; c. thca.1775</td>
<td>1831</td>
<td>1811</td>
<td>x</td>
<td>18th-19th</td>
<td>British</td>
</tr>
<tr>
<td>DANCE-HOLLAND, NATHANIEL</td>
<td>Dance-Holland, Nathaniel; Nathaniel</td>
<td>1735</td>
<td>1811</td>
<td>x</td>
<td>18th</td>
<td>British</td>
</tr>
</tbody>
</table>
and aligning to the Getty Vocabularies

Rembrandt van Rijn (Dutch painter, printmaker, 1606-1669)

Note: Noted as one of the most popular and influential artists of his period. His work is characterized by the Baroque interest in dramatic scenes and strong contrasts of light on a dark stage. The subjects of his works include paintings, etchings, and drawings of subjects including portraits, landscapes, figures, animals, biblical and secular history and mythology.

Names:
Rembrandt van Rijn (preferred, V, display, LC) .... "Rijn" refers to a geographic place, the site of the mill owned by his father in Leyden

Rijn, Rembrandt van (V)
Rembrandt Harmensz. van Rijn (V)
Rembrandt Harmensz van Rijn (V)
Rijn, Rembrandt Harmensz. van (V)
Van Rijn, Rembrandt (V)
Rembrandt (V, CN)

.... from 1632 onwards he signed his works with only the forename "Rembrandt"; in documents, however, he continued to sign "Rembrandt van Rijn"
and aligning to the Getty Vocabularies
Getting the vocabulary data
Turning it into JSON

-ensure that your target resource representations are consistent (e.g. no mixed types for a given field)

-include and dereference fields that will be most useful for matching (may require building dereferencing dictionaries to go from vocab IDs to string literals)

-ensure JSON is ES compatible (e.g. no blank node " " keys)
"personDescription": "Dutch painter, printmaker, 1606-1669",
"birthPlace": {
  "id": "tgn:7006809-place",
  "label": {
    "nl": ["Leiden", "Leyden"],
    "en": ["Leyden"]
  }
},
"deathPlace": {
  "id": "tgn:7006952-place",
  "label": {
    "nl": ["Amsterdam"],
    "en": ["Amsterdam"]
  }
},
"estStart": {
  "type": "xsd:Year",
  "@value": "1606-01-01"
},
"estEnd": {
  "type": "xsd:Year",
  "@value": "1669-01-01"
}
A few more steps

- Load JSON into your ES instance
- Optionally create a web service that can relay queries to the ES instance
- Consider how to validate reconciliation service
Querying the service

[
  {
    "_id": "500095103",
    "_index": "ulan",
    "_score": 6.190874,
    "_source": {
      "agentType": [
        {
          "id": "aat:300024987",
          "label": {
            "en": [
              "architects",
              "architect"
            ]
          }
        }
      ]
    }
  }
]
Validity: precision and recall

- Precision is your signal-to-noise ratio
  - Dense namespaces will reduce it
- Recall is the sensitivity to make matches
  - ES is very sensitive by default
- External validation will require labor
  - Known matches compared with results
One more idea
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>artist_authority_artist_1</td>
<td>variant_names_artist_1</td>
<td>birth_date_artist_1</td>
<td>death_date_artist_1</td>
<td>period_active_century_active_city_data_i</td>
<td>nationality_artist_1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ALLORI, ALESSANDRO</td>
<td>Allori, Alessandro; Alessandro Allori; alei</td>
<td>1535</td>
<td>1607</td>
<td>x</td>
<td>16th</td>
<td>Italian</td>
</tr>
<tr>
<td>3</td>
<td>ALLORI, ALESSANDRO</td>
<td>Allori, Alessandro; Alessandro Allori; alei</td>
<td>1535</td>
<td>1607</td>
<td>x</td>
<td>16th</td>
<td>Italian</td>
</tr>
<tr>
<td>4</td>
<td>MANLOL, ARISTIDE</td>
<td>Mallo, Aristide; Aristide Mallo; Mallo</td>
<td>1884</td>
<td>1944</td>
<td>x</td>
<td>19th-20th</td>
<td>French</td>
</tr>
<tr>
<td>5</td>
<td>COROT, JEAN BAPTISTE CAMILLE</td>
<td>Corot, Jean Baptiste Camille</td>
<td>1816</td>
<td>1877</td>
<td>x</td>
<td>19th</td>
<td>French</td>
</tr>
<tr>
<td>6</td>
<td>CHEMLINSKI, JAN</td>
<td>Chełmicki, Jacek; chełmicki</td>
<td>1891</td>
<td>1925</td>
<td>x</td>
<td>19th</td>
<td>Polish</td>
</tr>
<tr>
<td>7</td>
<td>PEAGO, CHARLES WILSON</td>
<td>Peake, Charles Wilson; Peale; peax c.w.</td>
<td>1741</td>
<td>1827</td>
<td>x</td>
<td>18th</td>
<td>American</td>
</tr>
<tr>
<td>8</td>
<td>PEAGO, CHARLES WILSON</td>
<td>Peake, Charles Wilson; Peale; peax c.w.</td>
<td>1741</td>
<td>1827</td>
<td>x</td>
<td>18th</td>
<td>American</td>
</tr>
<tr>
<td>9</td>
<td>BEMBRIDGE, HENRY</td>
<td>Bembridge, Henry; Henry Bembridge</td>
<td>1743</td>
<td>1812</td>
<td>x</td>
<td>18th</td>
<td>American</td>
</tr>
<tr>
<td>10</td>
<td>SULLY, THOMAS</td>
<td>Sulley, Thomas; Thomas Sully; Sulley</td>
<td>1733</td>
<td>1817</td>
<td>x</td>
<td>19th</td>
<td>American</td>
</tr>
<tr>
<td>11</td>
<td>CHASE, WILLIAM MERRITT</td>
<td>Chase, William Merritt; William Merritt</td>
<td>1849</td>
<td>1916</td>
<td>x</td>
<td>19th</td>
<td>American</td>
</tr>
<tr>
<td>12</td>
<td>REMBRANDT HARMENSZ. VAN Rijn</td>
<td>Rembrandt Harmensz. van Rijn; Rembra</td>
<td>1606</td>
<td>1669</td>
<td>x</td>
<td>17th</td>
<td>Amsterdam (1630); Dutch</td>
</tr>
<tr>
<td>13</td>
<td>BERTZIG, ANTON</td>
<td>Bertzig, Anton; Anton Bertz; bertzi</td>
<td>1856</td>
<td>1912</td>
<td>x</td>
<td>19th-20th</td>
<td>Austrian</td>
</tr>
<tr>
<td>14</td>
<td>TOMSON, CLIFTON</td>
<td>Tomson, Clifton; clifton tomson; c. thom ca.1775</td>
<td>1811</td>
<td>1811</td>
<td>x</td>
<td>18th-19th</td>
<td>British</td>
</tr>
<tr>
<td>15</td>
<td>DANCE-HOLLAND, NATHANIEL</td>
<td>Dance-Holland, Nathaniel; Nathaniel Dol</td>
<td>1735</td>
<td>1811</td>
<td>x</td>
<td>18th</td>
<td>London; Rome; British</td>
</tr>
<tr>
<td>16</td>
<td>SCOTT, SAMUEL</td>
<td>Scott, Samuel; Scott; Samuel Scott; Scott</td>
<td>1702/10</td>
<td>1773</td>
<td>x</td>
<td>18th</td>
<td>British</td>
</tr>
<tr>
<td>17</td>
<td>[ANONYMOUS - 18TH-C.]</td>
<td>[Anonymous - 18th-c.]; Anonymous - 18th-c.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>18th</td>
<td>Unknown</td>
</tr>
<tr>
<td>18</td>
<td>[ANONYMOUS - 18TH-C.]</td>
<td>[Anonymous - 18th-c.]; Anonymous - 18th-c.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>18th</td>
<td>Unknown</td>
</tr>
<tr>
<td>19</td>
<td>[ANONYMOUS - 18TH-C.]</td>
<td>[Anonymous - 18th-c.]; Anonymous - 18th-c.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>18th</td>
<td>Unknown</td>
</tr>
<tr>
<td>20</td>
<td>STEVENS, ALFRED EMILE LEOPOLDO JOSEPH VICTOR</td>
<td>Stevens, Alfred Emile Leopold Joseph Victor</td>
<td>1823</td>
<td>1906</td>
<td>x</td>
<td>19th</td>
<td>Belgian</td>
</tr>
<tr>
<td>21</td>
<td>MURPHY, JOHN FRANCIS</td>
<td>Murphy, John Francis; John Francis Mark</td>
<td>1853</td>
<td>1921</td>
<td>x</td>
<td>19th</td>
<td>American</td>
</tr>
<tr>
<td>22</td>
<td>WHISTLER, JAMES ABBOTT MCNEILL</td>
<td>Whistler, James Abbott McNeill; James A</td>
<td>1834</td>
<td>1903</td>
<td>x</td>
<td>19th</td>
<td>American</td>
</tr>
<tr>
<td>23</td>
<td>SÁNCHEZ PEREYRA, EMILIO</td>
<td>Sánchez Perieyra; Emilio; Emilio Sánchez</td>
<td>1815</td>
<td>1897</td>
<td>x</td>
<td>19th</td>
<td>Spanish</td>
</tr>
<tr>
<td>24</td>
<td>CHADWICK, LYNN RUSSELL</td>
<td>Chadwick, Lynn; Lynn chadwick</td>
<td>1914</td>
<td>2009</td>
<td>x</td>
<td>20th-21st</td>
<td>London; British</td>
</tr>
<tr>
<td>25</td>
<td>CHADWICK, LYNN RUSSELL</td>
<td>Chadwick, Lynn; Lynn chadwick</td>
<td>1914</td>
<td>2009</td>
<td>x</td>
<td>20th-21st</td>
<td>London; British</td>
</tr>
<tr>
<td>26</td>
<td>CHADWICK, LYNN RUSSELL</td>
<td>Chadwick, Lynn; Lynn chadwick</td>
<td>1914</td>
<td>2009</td>
<td>x</td>
<td>20th-21st</td>
<td>London; British</td>
</tr>
<tr>
<td>27</td>
<td>RAJULI, ETIENNE</td>
<td>Rajuli, Etienne; rajulu</td>
<td>1907</td>
<td>1996</td>
<td>x</td>
<td>20th</td>
<td>Paris; Switzerland; French</td>
</tr>
<tr>
<td>28</td>
<td>RAJULI, ETIENNE</td>
<td>Rajuli, Etienne; rajulu</td>
<td>1907</td>
<td>1996</td>
<td>x</td>
<td>20th</td>
<td>Paris; Switzerland; French</td>
</tr>
<tr>
<td>29</td>
<td>BRAQUE, GEORGES</td>
<td>Braque, Georges; Georges Braque; Bra</td>
<td>1882</td>
<td>1963</td>
<td>x</td>
<td>20th</td>
<td>French</td>
</tr>
<tr>
<td>30</td>
<td>POUGNY, JEAN</td>
<td>Pougny, Jean; jean Pougny; Pougny</td>
<td>1894</td>
<td>1996</td>
<td>x</td>
<td>20th</td>
<td>French</td>
</tr>
</tbody>
</table>
A review of the pieces

- vocabulary data represented in stand-alone JSON documents
- a loaded Elasticsearch instance
- a web service layer to relay queries to the ES instance (optional)
- source records to be reconciled
- a script for sending records to the reconciliation service
- a Google Drive account for storing and editing results
Reconciliation with Elasticsearch

Nathan Lomeli
nlomeli@getty.edu