Constructing a Focused Taxonomy from a Document Collection

Olena Medelyan, Steve Manion, Jeen Broekstra, and Anna Divoli

Anna Lan Huang and Ian Witten
Why Automatic Generation?
Dynamic
Fast
Cheap
Consistent
RDF / Flexible
...

Why from a Document Collection?
Focused/specific
Optimal for those documents
...
Talk Overview

The Team

The Process

Evaluation

News Group Case Study

Other Use Cases

Summary

@annadivoli
How Taxonomy Generation Works
**Input:**
Documents stored somewhere

**Analysis:**
Using variety of tools* and datasets, extract concepts, entities, relations

**Grouping & Output:**
An SKOS taxonomy is created that groups resulting taxonomy terms hierarchically
Taxonomy Generation in 5 Steps!

1. Import & convert to text
2. Extract concepts
3. Annotate with Linked Data
4. Disambiguate clashing concepts
5. Consolidate taxonomy

Input Docs

Document Database
Solr

Concepts & Relations Database
Sesame

Focused SKOS Taxonomy

Preferred top-level terms
Step 1. Document input & conversion

Current input:
- Directory path read recursively

Other possible inputs:
- Docs in a database or a DMS
- Emails + attachments (Exchange)
- Website URL
- RSS feed

External tool to convert different file formats to text

Database to store document content
http://localhost/solr/select?q=path:mycollection\document456.txt

"Ex-top climate official very concerned by global lack of progress"

DURBAN, South Africa - Yvo de Boer said he left his job as the U.N.'s top climate official in frustration 18 months ago, believing the process of negotiating a meaningful climate agreement is failing.

His opinion hasn't changed.

"I still have the same view of the process that led me to leave the process," he told the Associated Press on Sunday. "I'm still deeply concerned about where it's going, or rather where it's not going, about the lack of progress."

For three years until 2010, the Dutch civil servant was the leading voice on global warming on the world stage. He appeared constantly in public to advocate green policies, traveled endlessly for private meetings with top leaders and labored with negotiators seeking ways to finesse snags in drafting agreements.

In the end he felt he wasn't really able to contribute as I should be to the process," he said.

Today, he can take a long view on his years as a Dutch negotiator in the 1990s and later as a senior U.N. official with access to the highest levels of government, business and civil society. He is able to voice criticisms he was reluctant to air when he was actively shepherding climate diplomacy.

Negotiators live "in a separate universe," and the ongoing talks are "like a log that's drifted away," he said. Then, using another metaphor, he called the annual 194-nation conferences "a bit of a mouse wheel."

Elsewhere in Durban Sunday, the South African host of the talks called for divine help at a climate change church service organized by the South African Council of Churches.

"We needed to pray for [an] acceptable, balanced outcome, that has a sense of urgency," said Maite Nkoana-Mashabane, who as South Africa's foreign minister is president of the Durban round of negotiations.
Step 3. Annotation with meaning

Pingar API:
- People:
  - Yvo de Boer
  - Maite Nkoana-Mashabane
- Organizations:
  - Associated Press
  - South African Council of Churches
- Locations:
  - South Africa

mycollection/document456.txt
Step 3. Annotation with meaning

**Pingar API:**
- **People:** Yvo de Boer, Maite Nkoana-Mashabane
- **Organizations:** Associated Press, South African Council of Churches
- **Locations:** South Africa

Later this additional info will help create e-Discovery & semantic search solutions.

```
mycollection/document456.txt
```
Final Concepts

4. Disambiguate clashing concepts

Final Concepts

Over the past three years, Apple has acquired three mapping companies.

Two concepts were extracted, that are **dissimilar**
Discard the incorrect one

Apple (Apple Corps) ≠ Apple Inc.

wikipedia.org/wiki/Apple_Corps freebase.com/view/en/apple_inc

For millions of years, the oceans have been filled with sounds from natural sources.

Two concepts were extracted, that are **similar**
Accept both correct

Agrovoc term: Marine areas

wikipedia.org/wiki/Ocean www.fao.org/aos/agrovoc#c_4607
Step 5a. Group taxonomy

Building the taxonomy bottom up
Step 5a. Group taxonomy

Concepts & Relations Database → 5a. Add relations → Focused SKOS Taxonomy

Building the taxonomy bottom up

Felines: zebra, donkey, tiger
Lizard, horse, bird, pigeon
Step 5a. Group taxonomy

Building the taxonomy bottom up
Step 5a. Group taxonomy

Building the taxonomy bottom up

Concepts & Relations Database

5a. Add relations

Focused SKOS Taxonomy

animals

horse family

felines  zebra  donkey  tiger  lizard  horse  bird  pigeon

Category: Carnivorous animals

Category: Animals
Step 5a. Group taxonomy

Concepts & Relations Database ➔ 5a. Add relations ➔ Focused SKOS Taxonomy

Animals

- Felines
- Zebra
- Donkey
- Tiger
- Lizard
- Horse
- Bird
- Pigeon

Category: Carnivorous animals

Broader: Squamata/Reptiles/Tetrapods/Vertebrates/Chordates/Animals

Building the taxonomy bottom up
Step 5b. Consolidating taxonomy

Concepts & Relations Database → 5b. Prune relations → Focused SKOS Taxonomy

Films and film making
- Film stars
  - Mila Kunis
  - Daniel Radcliffe
  - Sally Hawkins
  - Julianna Margulies

Association football clubs
- Former Football League clubs
  - Manchester United F.C.
  - Manchester City F.C.

Finance
- Economics and finance
- Personal finance
- Commercial finance
- Tax
  - Capital gains tax

5b. Prune relations
Vocabulary of Ngrams, Concepts and Entities shared across various tools. All intermediate processing data is captured and stored using RDF triples. The data can be queried using the SPARQL query language.
**Taxonomy Generation Process**

**Input:**
Documents stored somewhere

**Analysis:**
Using variety of tools* and datasets, extract concepts, entities, relations

**Output:**
An SKOS taxonomy is created that groups resulting taxonomy terms hierarchically

* Pingar API for People, Organization, Locations & Taxonomy Terms from related taxonomies; Wikification for related Wikipedia articles and category relations; Linked Data analysis for creating links to Freebase & DBpedia
Case Study & Evaluation: A News Group

Fairfax NZ
This taxonomy was created from 2000 news articles by Fairfax New Zealand around Christmas 2011. (4.3MB of uncompressed text, averaging ~ 300 words each)
+ UK Integrated Public Service Sector vocabulary (http://doc.esd.org.uk/IPSV/2.00.html)

Taxonomy Statistics
Concept Count: 10158
Edges Count: 12668
Intermediate Count: 1383
Leaves Count: 8748
Labels Count: 11545

Nesting Counts
0: 27, 1: 6102, 2: 2903, 3: 2891
4: 2057, 5: 1202, 6: 745, 7: 354
8: 179, 9: 41, 10: 10

Average Depth: 2.65
Evaluation
Evaluation

Coverage: 75%

Comparing with manually generated taxonomy by Fairfax librarians for the same domain (458 concepts - was never completed).

Some not really missing: “Drunk” vs. “Drinking alcohol” and “Alcohol use and abuse”
Truly missing: “Immigration”, “Laptop” and “Hospitality”
Evaluation

**Coverage: 75%**

Comparing with manually generated taxonomy by Fairfax librarians for the same domain (458 concepts - was never completed).

Some not really missing: “Drunk” vs. “Drinking alcohol” and “Alcohol use and abuse”

Truly missing: “Immigration”, “Laptop” and “Hospitality”

**Precision** (15 human judges based evaluation):

**90% for relations**

100 concept pairs - yes/no decision whether relation makes sense. Total of 750 relations examined – each by two different judges.

Examples: “North Yorkshire → Leeds”, “Israel → History of Israel”

Humans: “Infectious Disease → Polio”, “Scandinavia → Sweden” !

**89% for concepts...**
Evaluation: Sources of error in concept identification
Evaluation: Sources of error in concept identification

... Precision (15 human judges based evaluation):

89% for concepts
Given extracted concepts and original text.
300 documents equally distributed plus 5 to all judges.
### Evaluation: Sources of error in concept identification

**... Precision** (15 human judges based evaluation):

89% for concepts

Given extracted concepts and original text.
300 documents equally distributed plus 5 to all judges.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Errors</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>1145</td>
<td>37</td>
<td>3.2%</td>
</tr>
<tr>
<td>Organizations</td>
<td>496</td>
<td>51</td>
<td>10.3%</td>
</tr>
<tr>
<td>Locations</td>
<td>988</td>
<td>114</td>
<td>11.5%</td>
</tr>
<tr>
<td>Wikipedia named entities</td>
<td>832</td>
<td>71</td>
<td>8.5%</td>
</tr>
<tr>
<td>Wikipedia other entities</td>
<td>99</td>
<td>16</td>
<td>16.4%</td>
</tr>
<tr>
<td><strong>Taxonomy</strong></td>
<td>868</td>
<td>229</td>
<td>26.4%</td>
</tr>
<tr>
<td>DBPedia</td>
<td>868</td>
<td>81</td>
<td>8.1%</td>
</tr>
<tr>
<td>Freebase</td>
<td>135</td>
<td>12</td>
<td>8.9%</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>3447</td>
<td>393</td>
<td>11.4%</td>
</tr>
</tbody>
</table>
Case Study: A News Group

- National News (588)
  - Agriculture (90)
  - Catteries
  - Climate and weather (18)
  - Education (43)
  - Environment (363)
  - Flowers (12)
  - Seasons (44)
  - Shops (2)
  - Stables
  - Surveying (2)
  - Warehousing
  - Workshops (1)
  - Zoning

- Public order, justice and rights (169)
  - Arrest
  - Civil and human rights
  - Court orders (1)
  - Crime (29)
  - Emergencies (26)
  - Fraud (2)
  - Hearings
  - Judiciary (11)
  - Law (71)
  - Legislation (1)
  - Offenders (4)
  - Prisons (8)
  - Prosecutions
  - Remand
  - Sentencing
  - Witnesses
Case Study: A News Group

- Business (499)
  - Education and skills (114)
    - Further and higher education (15)
      - Lecturers (1)
      - Postgraduate education
    - University (11)
      - Canterbury University
      - Charles University in Prague
      - London University
      - New York University
      - Otago University
      - Oxford University
      - Pontifical Urbanian University
      - Prague University
      - Stanford University
      - Victoria University
      - Waikato University
  - Schools (94)
  - Skills and competences (2)
- Entertainment (173)
- Food and drink (398)
- Government, politics and public administration (222)
- Health, well-being and care (235)
- Housing (5)
- Information and communication (601)
- International affairs and defence (79)
- Leisure and culture (568)
- Life in the community (419)
- Locations (3346)
- National News (588)

- Health, well-being and care (235)
  - Benefits
  - Care (9)
  - Diabetes (2)
  - Disability (2)
  - Explosives (1)
  - Health (166)
  - Hospitals (28)
    - Anderson's Veterinary Hospital
    - Auckland Hospital
    - Burwood Hospital
    - Children's hospital
    - Christchurch Charity Hospital
    - Christchurch Hospital
    - Dunedin Hospital
    - Gold Coast Hospital
    - Hospital
    - Hospital departments (2)
      - Kalkoura Hospital
      - Lady Duncan Hospital
      - Matamata Chronicle Hospital
      - Middlemore Hospital
      - Pioneering Polio Hospital
      - Pittsburgh Children's Hospital
      - Princess Margaret Hospital
      - Santa Casa de Misericordia Hospital
      - Southland Hospital
      - Southland Hospital
      - Stratford Hospital
      - Taranaki Base Hospital
      - Timaru Hospital
      - Whakatane Hospital
      - Wellington Hospital
      - Waikato Hospital
## Case Study: A News Group

<table>
<thead>
<tr>
<th>Topic</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>People (5023)</td>
<td></td>
</tr>
<tr>
<td>Public order, justice and rights (169)</td>
<td></td>
</tr>
<tr>
<td>Royal Christmas Message</td>
<td></td>
</tr>
<tr>
<td>Science, technology and innovation (560)</td>
<td></td>
</tr>
<tr>
<td>Astronomy and space (3)</td>
<td></td>
</tr>
<tr>
<td>Comet (2)</td>
<td></td>
</tr>
<tr>
<td>C/2006 P1</td>
<td></td>
</tr>
<tr>
<td>Sungrazing comet</td>
<td></td>
</tr>
<tr>
<td>Biology and applied life sciences (44)</td>
<td></td>
</tr>
<tr>
<td>Biotechnology (1)</td>
<td></td>
</tr>
<tr>
<td>Vaccination</td>
<td></td>
</tr>
<tr>
<td>Engineering (51)</td>
<td></td>
</tr>
<tr>
<td>Environmental and earth sciences (93)</td>
<td></td>
</tr>
<tr>
<td>Mathematics (2)</td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td></td>
</tr>
<tr>
<td>Statistical analysis</td>
<td></td>
</tr>
<tr>
<td>Physics (24)</td>
<td></td>
</tr>
<tr>
<td>Research and development (2)</td>
<td></td>
</tr>
<tr>
<td>Research (1)</td>
<td></td>
</tr>
<tr>
<td>Atmospheric Research</td>
<td></td>
</tr>
<tr>
<td>Science (263)</td>
<td></td>
</tr>
<tr>
<td>Technology (67)</td>
<td></td>
</tr>
<tr>
<td>Science, technology and innovation (560)</td>
<td></td>
</tr>
<tr>
<td>Single Nodes (554)</td>
<td></td>
</tr>
<tr>
<td>Sports and recreation (416)</td>
<td></td>
</tr>
<tr>
<td>Aeronautical activities (2)</td>
<td></td>
</tr>
<tr>
<td>Athletics (11)</td>
<td></td>
</tr>
<tr>
<td>Bowling sports and games (1)</td>
<td></td>
</tr>
<tr>
<td>Cooling down</td>
<td></td>
</tr>
<tr>
<td>Crossword (1)</td>
<td></td>
</tr>
<tr>
<td>Field sports (2)</td>
<td></td>
</tr>
<tr>
<td>Football (1)</td>
<td></td>
</tr>
<tr>
<td>Golf</td>
<td></td>
</tr>
<tr>
<td>Horse racing (12)</td>
<td></td>
</tr>
<tr>
<td>League (111)</td>
<td></td>
</tr>
<tr>
<td>Martial arts (3)</td>
<td></td>
</tr>
<tr>
<td>Motor sports (4)</td>
<td></td>
</tr>
<tr>
<td>New Zealand national rugby union team (1)</td>
<td></td>
</tr>
<tr>
<td>Outdoor pursuits (11)</td>
<td></td>
</tr>
<tr>
<td>Pet</td>
<td></td>
</tr>
<tr>
<td>Premier League (1)</td>
<td></td>
</tr>
<tr>
<td>Professional golfer (15)</td>
<td></td>
</tr>
<tr>
<td>Racing (16)</td>
<td></td>
</tr>
<tr>
<td>Rugby (7)</td>
<td></td>
</tr>
<tr>
<td>Snowman</td>
<td></td>
</tr>
<tr>
<td>Team sports (161)</td>
<td></td>
</tr>
<tr>
<td>Tennis (9)</td>
<td></td>
</tr>
<tr>
<td>Water sports (21)</td>
<td></td>
</tr>
<tr>
<td>Winter sports (2)</td>
<td></td>
</tr>
<tr>
<td>Transport and infrastructure (206)</td>
<td></td>
</tr>
</tbody>
</table>
Alternative Labels

Counselling

Concept ID: http://pingar.com/waikato/sample2000/docs_Dec11_xmas-Counselling

Preferred Label

Counselling

Alternative Labels

- Counseling
Marriage and cohabitation

Preferred Label

Marriage and cohabitation

Alternative Labels

- Cohabitation
- Husband
- Living together
- Marriage
- Matrimony
- Wife
### Labels & Relations

#### Primary schools

<table>
<thead>
<tr>
<th>Labels</th>
<th>Relations</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preferred Label</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary schools</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alternative Labels</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First schools (primary schools)</td>
<td></td>
</tr>
<tr>
<td>Primary education</td>
<td></td>
</tr>
</tbody>
</table>

**Concept ID:** http://pingar.com/waikato/fairfaxSep01-2-Primary%20schools

#### Broader Relations

- School types

#### Narrower Relations

- Takaka Primary School
- Wainoni Primary School

#### Related Relations
Case Study: A News Group

Fairfax - 4 Days from Sep 2001

Excerpt of the taxonomy generated from:
Fairfax articles taken from
- Sep 9th & 10th (1242 articles) and
- Sep 13th & 14th (1667 articles) NZT!

Colors of terms:
- **proposed** to group other terms
- found in **both** document collections
- in **9-10 Sep 2001** docs
- in **13-14 Sep 2001** docs
- search match

**Taxonomy Statistics:**
Concept Count: 12699
Edges Count: 13755
Intermediate Count: 709
Leaves Count: 11985
Labels Count: 12741
Case Study: A News Group

proposed to group other terms
in both document collections
in 9-10 Sep 2001 docs
in 13-14 Sep 2001 docs

B Australia
B Blair
B Clark
B Clive
Billy Wilder
Bilton
Bin Laden
Bing Lucas
Bingham Group
Binyamin Hochner
Bio Dynamic Association
Biodiversity
Biology and applied life sciences
Biotechnology

Osama
Osama Abusidou
Osama ben Laden
Osbourne Ruddock
Oscar
Case Study: A News Group

proposed to group other terms in both document collections in 9-10 Sep 2001 docs in 13-14 Sep 2001 docs
Case Study: A News Group

September 2001

- Science, technology and innovation (11)
  - Environmental and earth sciences (5)
    - Geography (10)
      - Continents (3)
        - Europe (1)
        - Eastern Europe (2)
        - Turkey

- Island (23)
  - Christmas Island

proposed to group other terms in both document collections in 9-10 Sep 2001 docs in 13-14 Sep 2001 docs

Christmas 2011

- Terror
  - Life in the community (15)
    - Abuse (13)
      - Terrorism (5)
  - Public order, justice and rights (16)
    - Crime (25)
      - Organised crime and terrorism
  - Single Nodes (109)
    - Islamic politics and Islamic world studies (1)
      - Islamic fundamentalism (6)
        - Al-Qaeda (1)
        - War on Terror (4)
Other Use Cases

What’s in these files / emails?

How to refine search by metadata?

What to include into our corporate taxonomy?

How to find all docs on a given topic?

Content Audit

Better search with facets

Information Architecture

Better browsing
proposed to group other concepts in two or more document collections in the **bipolar** document collection in the **breast cancer** document collection in the **neither** cancer or bipolar doc. collection.
Summary

- Entity Extraction
- Linked Data
- Disambiguation
- Consolidation
- Evaluation
- News Group Case Study
- Other Use Cases
Focused SKOS Taxonomy Extraction Process (F-STEP) wiki

F-STEP Taxonomies

Home

This website showcases taxonomies generated by a taxonomy generator, called F-STEP.

F-STEP, which stands for Focused SKOS Taxonomy Extraction Process, automatically creates taxonomies focused on a particular document collection by utilizing public external sources that define the semantics of relevant taxonomy concepts, including Freebase, DBPedia, Wikipedia, and, if available, additional thesauri or taxonomies for related domains, e.g. IPSY or Agrovoc. The result is a custom-built taxonomy, expressed in SKOS, a standard RDF format.

Taxonomy Generation: Overall Process

Input: Documents stored somewhere

Analysis: Using variety of tools* and datasets, extract concepts, entities & relations

Output: A taxonomy is created that groups resulting taxonomy terms hierarchically

*Google, Freebase, DBPedia, Wikipedia, etc.

More?

@PingarHQ

anna.divoli@pingar.com

@annadivoli

bit.ly/f-step
Additional Slides
Query results in the Sesame workbench using the output generated during Taxonomy Generation

<table>
<thead>
<tr>
<th>Name</th>
<th>BdDate</th>
<th>Gender</th>
<th>Doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan Hunt’en</td>
<td>1927-10-09</td>
<td>male</td>
<td><a href="http://devwikiminer.pinger.net/solr/select?q-path:sample2000docs-Dec11_xmas%5Cs0485396.txt">http://devwikiminer.pinger.net/solr/select?q-path:sample2000docs-Dec11_xmas%5Cs0485396.txt</a></td>
</tr>
<tr>
<td>Bill Clinton’en</td>
<td>1946-08-19</td>
<td>male</td>
<td><a href="http://devwikiminer.pinger.net/solr/select?q-path:sample2000docs-Dec11_xmas%5Cs048158.txt">http://devwikiminer.pinger.net/solr/select?q-path:sample2000docs-Dec11_xmas%5Cs048158.txt</a></td>
</tr>
<tr>
<td>Bill Clinton’en</td>
<td>1946-08-19</td>
<td>male</td>
<td><a href="http://devwikiminer.pinger.net/solr/select?q-path:sample2000docs-Dec11_xmas%5Cs0486117.txt">http://devwikiminer.pinger.net/solr/select?q-path:sample2000docs-Dec11_xmas%5Cs0486117.txt</a></td>
</tr>
</tbody>
</table>
The Format of the Exported* Taxonomy

* We also support export into SharePoint Term Store format