

# **Extracting and Formatting Patent Data from USPTO XML**

Gabe Fierro

# College of Engineering University of California, Berkeley

Fung Technical Report No. 2013.06.10 http://www.funginstitute.berkeley.edu/sites/default/files/Extracting&Formatting\_Patent Data.pdf

June 10, 2013

The Coleman Fung Institute for Engineering Leadership, launched in January 2010, prepares engineers and scientists – from students to seasoned professionals – with the multidisciplinary skills to lead enterprises of all scales, in industry, government and the nonprofit sector.

Headquartered in UC Berkeley's College of Engineering and built on the foundation laid by the College's Center for Entrepreneurship & Technology, the Fung Institute combines leadership coursework in technology innovation and management with intensive study in an area of industry specialization. This integrated knowledge cultivates leaders who can make insightful decisions with the confidence that comes from a synthesized understanding of technological, marketplace and operational implications.

Copyright © 2013, by the author(s). All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

### **Lee Fleming,** Faculty Director, Fung Institute

# **Advisory Board**

#### **Coleman Fung**

Founder and Chairman, OpenLink Financial

#### **Charles Giancarlo**

Managing Director, Silver Lake Partners

#### **Donald R. Proctor**

Senior Vice President, Office of the Chairman and CEO, Cisco

#### In Sik Rhee

General Partner, Rembrandt Venture Partners

## **Fung Management**

# **Lee Fleming**

**Faculty Director** 

#### Ikhlaq Sidhu

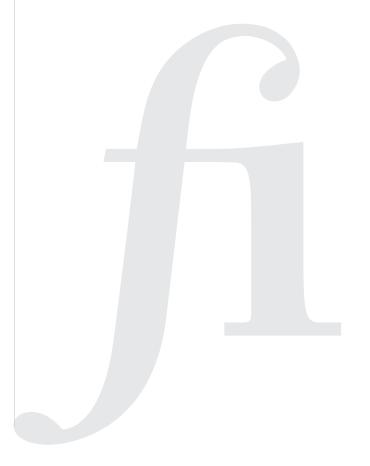
Chief Scientist and CET Faculty Director

#### **Robert Gleeson**

**Executive Director** 

#### **Ken Singer**

Managing Director, CET



**Abstract:** I describe data formatting problems that arise from extracting useful and relevant data from the XML files distributed by USPTO. I then describe solutions for a consistent data schematic that dictates in what format the extracted data fields should be stored and how these transformations should be applied to the data.