A Pattern Recognition and NLP Approach to Automated XBRL Extraction

Hassan Alam, Aman Kumar, Cheryl Lee, Yuliya Tarnikova

BCL Technologies
San Jose, CA 95128
Introduction

- XBRL (eXtensible Business Reporting Language) is a technical standard for describing financial and related data
- It’s a way to assign standard tags to financial data in reports and systems
- BCL Technologies has developed a simplified and automated way to create an SEC-compliant XBRL document
- In this study, we developed an automated system to derive text and tables from 10Q and 10K financial documents and do detail tagging
Flowchart of the SmartXBRL system

1. **Data Extraction**
   - Company Financial Source Documents
   - HTML

2. **XBRL Generation**
   - Data Files
   - Seq | Description | Document
   - 14 | XBRL INSTANCE DOCUMENT | aa-20111231.xml
   - 15 | XBRL TAXONOMY EXTENSION SCHEMA | aa-20111231.xsd
   - 16 | XBRL TAXONOMY EXTENSION LINKBASE | aa-20111231_def.xml
   - 17 | XBRL TAXONOMY EXTENSION LABEL LINKBASE | aa-20111231_lab.xml
   - 18 | XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE | aa-20111231_pre.xml

3. **Syntactic-Semantic Tagging**
   - XBRL Tagged Data

4. **Data Verification**
   - Government Filing

5. **XBRL Generation**
Flowchart of the SmartXBRL system

1. **Company Financial Source Documents**
   - Data Extraction
   - Syntactic-Semantic Tagging
     - XBRL Tagged Data
   - XBRL Generation
     - Government Filing
   - Data Verification

Diagram flow:
- Company Financial Source Documents
- Data Extraction
- Syntactic-Semantic Tagging
- XBRL Tagged Data
- XBRL Generation
- Government Filing
- Data Verification
Flowchart of the SmartXBRL system

Data Verification

Company Financial Source Documents

Data Extraction

Syntactic-Semantic Tagging

XBRL Tagged Data

Government Filing

XBRL Generation

Data Verification

Company Financial Source Documents

Data Extraction

Syntactic-Semantic Tagging

XBRL Tagged Data

Government Filing

XBRL Generation

Data Verification
Primary Goals of this Study

- Identify and extract DEI (Document Entity Information)
- Locate, extract, and modify the four financial tables (Income statement, Cash Flow, Balance Sheet, and Stockholders’ Equity)
- Extract and present parenthetical information from the four financials in a tabular format
- Extract financial Notes
- Detail tag
A typical 10Q document has three main parts:

- **Document Entity Information**
- **Part I** – financial information that includes financial tables and Notes, ‘Management’s Discussion and Analysis of Financial Condition and Results of Operations’ ETC.
- **Part II**, broadly speaking, includes sections on legal proceedings, risk factors ETC.
United States
Securities and Exchange Commission
Washington, D.C. 20549

Form 10-Q

(Dated One)

Quarterly Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934

For the Quarterly Period Ended September 30, 2011

Or

Transition Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934

Commission File Number 1-3610

<table>
<thead>
<tr>
<th></th>
<th>Third quarter ended September 20, 2011</th>
<th>Third quarter ended September 29, 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales (F)</td>
<td>$6,419</td>
<td>$5,287</td>
</tr>
<tr>
<td>Cost of goods sold (exclusive of expenses below)</td>
<td>5,290</td>
<td>4,413</td>
</tr>
<tr>
<td>Selling, general administrative, and other expenses</td>
<td>261</td>
<td>232</td>
</tr>
<tr>
<td>Research and development expenses</td>
<td>47</td>
<td>40</td>
</tr>
<tr>
<td>Provision for depreciation, depletion, and amortization</td>
<td>216</td>
<td>216</td>
</tr>
<tr>
<td>Restructuring and other charges (D)</td>
<td>9</td>
<td>49</td>
</tr>
<tr>
<td>Interest expense</td>
<td>123</td>
<td>139</td>
</tr>
<tr>
<td>Other expenses (income), net (F)</td>
<td>31</td>
<td>43</td>
</tr>
<tr>
<td>Total costs and expenses</td>
<td>6,139</td>
<td>5,227</td>
</tr>
<tr>
<td>Income from continuing operations before income taxes</td>
<td>280</td>
<td>60</td>
</tr>
<tr>
<td>Provision (benefit) for income taxes (D)</td>
<td>55</td>
<td>(40)</td>
</tr>
</tbody>
</table>

Notes to the Consolidated Financial Statements (unaudited)
(dollars in millions, except per-share amounts)

A. Basis of Presentation – The interim Consolidated Financial Statements of Alcoa Inc. and its subsidiaries ("Alcoa" or the "Company") are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company's results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa's Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.

B. Recently Adopted and Recently Issued Accounting Guidance
**Sample 10Q**

**UNITED STATES SECURITIES AND EXCHANGE COMMISSION**  
WASHINGTON, D.C. 20549

**FORM 10-Q**

☐ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the Quarterly Period Ended September 30, 2011

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

Commission File Number 1-3610

<table>
<thead>
<tr>
<th>Third quarter ended September 29,</th>
<th>Nine months ended September 29,</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>2012</td>
</tr>
<tr>
<td>Sales ($)</td>
<td>$8,419</td>
</tr>
<tr>
<td>Cost of goods sold (exclusive of expenses below)</td>
<td>5,290</td>
</tr>
<tr>
<td>Selling, general administrative, and other expenses</td>
<td>261</td>
</tr>
<tr>
<td>Research and development expenses</td>
<td>47</td>
</tr>
<tr>
<td>Provision for depreciation, depletion, and amortization</td>
<td>376</td>
</tr>
<tr>
<td>Restructuring and other charges (D)</td>
<td>9</td>
</tr>
<tr>
<td>Interest expense</td>
<td>123</td>
</tr>
<tr>
<td>Other expenses (income), net ($)</td>
<td>31</td>
</tr>
<tr>
<td>Total costs and expenses</td>
<td>6,139</td>
</tr>
<tr>
<td>Income from continuing operations before income taxes</td>
<td>280</td>
</tr>
<tr>
<td>Provision (benefit) for income taxes (M)</td>
<td>55</td>
</tr>
</tbody>
</table>

Notes to the Consolidated Financial Statements (unaudited)  
(dollars in millions, except per-share amounts)

A. Basis of Presentation — The interim Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (“Alcoa” or the “Company”) are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company’s results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa’s Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.

B. Recently Adopted and Recently Issued Accounting Guidance
UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
WASHINGTON, D.C. 20549

FORM 10-Q

(Oct One)

☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES
EXCHANGE ACT OF 1934

For the Quarterly Period Ended September 30, 2011

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES
EXCHANGE ACT OF 1934

Commission File Number 1-3610

 Third quarter ended Nine months ended
September 30, September 30, September 30, September 30,
2011 2010 2011 2010
Sales ($) 6,419 3,287 18,962 15,361
Cost of goods sold (exclusive of expenses below) 5,290 4,413 15,252 12,636
Selling, general administrative, and other expenses 261 232 759 679
Research and development expenses 47 40 136 134
Provision for depreciation, depletion, and amortization 376 338 1,112 1,079
Restructuring and other charges ($D) 96 2 49 219
Interest expense 123 139 399 376
Other expenses (income), net ($) 31 43 (47) 48
Total costs and expenses 6,139 3,227 17,600 15,161
Income from continuing operations before income taxes 280 60 1,302 200
Provision (benefit) for income taxes ($M) 55 (40) 328 92

Notes to the Consolidated Financial Statements (unaudited)
(dollars in millions, except per-share amounts)

A. Basis of Presentation – The interim Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (“Alcoa” or the “Company”) are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company’s results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa’s Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.

B. Recently Adopted and Recently Issued Accounting Guidance
Sample 10Q

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
WASHINGTON, D.C. 20549

FORM 10-Q

☐ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES
EXCHANGE ACT OF 1934

For the Quarterly Period Ended September 30, 2011

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES
EXCHANGE ACT OF 1934

Commission File Number 1-3610

<table>
<thead>
<tr>
<th>Third quarter ended September 30,</th>
<th>Nine months ended September 30,</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>2010</td>
</tr>
<tr>
<td>Sales (1)</td>
<td>$6,419</td>
</tr>
<tr>
<td>Cost of goods sold (exclusive of expenses below)</td>
<td>5,290</td>
</tr>
<tr>
<td>Selling, general administrative, and other expenses</td>
<td>261</td>
</tr>
<tr>
<td>Research and development expenses</td>
<td>47</td>
</tr>
<tr>
<td>Provision for depreciation, depletion, and amortization</td>
<td>376</td>
</tr>
<tr>
<td>Restructuring and other charges (2)</td>
<td>9</td>
</tr>
<tr>
<td>Interest expense</td>
<td>123</td>
</tr>
<tr>
<td>Other expenses (income), net (1)</td>
<td>31</td>
</tr>
<tr>
<td>Total costs and expenses</td>
<td>6,139</td>
</tr>
<tr>
<td>Income from continuing operations before income taxes</td>
<td>280</td>
</tr>
<tr>
<td>Provision (benefit) for income taxes (M)</td>
<td>55</td>
</tr>
</tbody>
</table>

Notes to the Consolidated Financial Statements (unaudited)
(dollars in millions, except per-share amounts)

A. Basis of Presentation — The interim Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (“Alcoa” or the “Company”) are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company’s results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa’s Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.

B. Recently Adopted and Recently Issued Accounting Guidance
DEI Extraction

A sample output of BCL’s SmartXBRL DEI extraction
DEI Extraction

- In a 10Q document the DEI information is usually given in the beginning part of the document.
- A heuristic-based pattern recognition algorithm identifies the DEI information.
- Example: for ‘Document Type,’ the keyword to look for is `<Form>` followed by `<document type::10-Q>`
Accuracy is around 89% (number of correct DEI fields extracted for 10 companies)

DEI is edited manually after the extraction to account for 100% accuracy for filing purposes
FINANCIAL TABLE EXTRACTION

Tasks

- Locate a financial table in a 10Q document
- Identify and differentiate a table layout from normal text, images, bulleted points, etc.
- Identify the header lines and differentiate between a column or row header and the title in a table
- Link the cell content to the cell tag (header, title, etc.)
Heuristic-based methods for extracting table information by utilizing the absolute position of text elements in a document

First the document is segmented and analyzed by means of image processing techniques in order to detect potential table areas

Parts are then passed to an OCR engine and text processing module that analyzes the geometry, syntax, and the semantics of the character data
Previous Work Cont'd.

- Scanned image
  - Table Isolation
    - Connected Component Analysis
    - OCR
  - Syntactic Analysis
  - Semantic Analysis
  - Table Components
In this work we are able to:

- Isolate embedded tables from documents
- Identify table components such as title blocks, table entries, and footer blocks
<table>
<thead>
<tr>
<th>Number of Samples</th>
<th>Detection of Tables</th>
<th>Detection of Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Scientific Journals</td>
<td>100</td>
<td>0.930</td>
</tr>
<tr>
<td>Financial Tables</td>
<td>100</td>
<td>0.714</td>
</tr>
</tbody>
</table>
The algorithm is geared toward finding a minimum set of rectangles that covers all white space of a page.

Tables are identified by assuming that text and images are all represented by a bounding rectangle with sides parallel to page borders.
Extracting Four Face Financial Tables

Pre-processing
- stop words, monetary, financial and temporal terms

Pattern Matching
- Concept Search
- Template Pattern Search

- concept identification
- semantic processing

Template Generation
- output: ‘cash flow’
Extracting Four Face Financial Tables

- stop words, monetary, financial and temporal terms

- concept identification
- semantic processing

- output: ‘cash flow’
Extracting Four Face Financial Tables

- Pre-processing
  - stop words, monetary, financial and temporal terms

- Pattern Matching
  - Concept Search
  - Template Pattern Search

- Concept identification
- Semantic processing

- Template Generation
  - output: ‘cash flow’
Extracting Four Face Financial Tables

- Pre-processing
  - stop words, monetary, financial and temporal terms

- Pattern Matching
  - Concept Search
  - Template Pattern Search
    - concept identification
    - semantic processing

- Template Generation
  - output: ‘cash flow’
Parenthetical tables are built when the line item in a face financial table discloses an amount in the input financial document.

The task is to extract the line items from the financial tables, which disclose parenthetical information and build a separate table.
### GENERATE AUTO PARENTHETICALS

<table>
<thead>
<tr>
<th>(in thousands)</th>
<th>2007</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receivables, less allowances of $1,260 and $1,150</td>
<td>31,659</td>
<td>31,601</td>
</tr>
</tbody>
</table>

- String matching algorithm
- Natural Language Processing methods

<table>
<thead>
<tr>
<th>(in thousands)</th>
<th>2007</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receivables, less allowances</td>
<td>31,659</td>
<td>31,601</td>
</tr>
<tr>
<td>(Less allowances)</td>
<td>(1,260)</td>
<td>(1,150)</td>
</tr>
</tbody>
</table>
LOCATE AND EXTRACT NOTES

- Notes are generally added to the financial documents to explain items on the financial statements.
- For Year 1 filers’, Notes are expected to be tagged as ‘block’ text.
- The identification and extraction of the Notes is done based on pattern-matching algorithm that also looks into the typological position of the notes block in a document.
Snapshot of Notes output by SmartXBRL

<table>
<thead>
<tr>
<th>Basis of Presentation - The interim</th>
<th>2010-09-30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basis of Presentation - The interim Cons:</td>
<td>2011-06-30</td>
</tr>
</tbody>
</table>

Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (Alcoa 8220; Company 8221;) are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company’s results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa’s Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.
**Basis of Presentation - The interim**

<table>
<thead>
<tr>
<th>Year</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-09-01</td>
<td></td>
</tr>
<tr>
<td>2011-06-30</td>
<td></td>
</tr>
</tbody>
</table>

Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (including Alcoa & 8221; or the & 8220;Company& 8221;) are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company's results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa's Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.
**Snapshot of Notes output by SmartXBRL**

![SmartXBRL 1.0 interface](image)

| Note | Note 1 | Note 2 | Note 3 | Note 4 | Note 5 | Note 6 | Note 7 | Note 8 | Note 9 | Note 10 | Note 11 | Note 12 | Note 13 | Note 14 | Note 15 | Note 16 | **Note 17** |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|

### Add Abstract Row

**Basis of Presentation - The interim**

- 2010-09-01
- 2011-06-30

**Basis of Presentation - The interim Cons:**

---

Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (Alcoa® and the Aluminum Company®) are unaudited. These Consolidated Financial Statements include all adjustments, consisting of normal recurring adjustments, considered necessary by management to fairly state the Company’s results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa®’s Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.
Snapshot of Notes output by SmartXBRL

SmartXBRL 1.0

[Add Abstract Row]

<table>
<thead>
<tr>
<th>Basis of Presentation - The interim</th>
<th>2010-09-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-06-30</td>
<td></td>
</tr>
</tbody>
</table>

Basis of Presentation - The interim Cons:
A. Basis of Presentation & Indicative results of operations, financial position, and cash flows. The results reported in these Consolidated Financial Statements are not necessarily indicative of the results that may be expected for the entire year. The 2010 year-end balance sheet data was derived from audited financial statements but does not include all disclosures required by accounting principles generally accepted in the United States of America (GAAP). This Form 10-Q report should be read in conjunction with Alcoa's Annual Report on Form 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.
At January 1, 2011 and January 2, 2010, there were cash overdrafts of approximately $9.0 million and $11.6 million, respectively, which are included in accounts payable within the accompanying consolidated balance sheets.

<table>
<thead>
<tr>
<th>Consolidated Balance Sheets (USD)</th>
<th>Jan. 01, 2011</th>
<th>Jan. 02, 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounts Payable</td>
<td>$9.0</td>
<td>$11.6</td>
</tr>
<tr>
<td>Cash Overdraft</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# RESULTS of XBRL Data Extraction

<table>
<thead>
<tr>
<th></th>
<th>Number of 10-Q Samples</th>
<th>Detection</th>
<th>Extraction of Cells/Content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Tables</td>
<td>150</td>
<td>.96</td>
<td>.95</td>
</tr>
<tr>
<td>DEI</td>
<td>150</td>
<td>.97</td>
<td>.95</td>
</tr>
<tr>
<td>Parenthetical</td>
<td>150</td>
<td>.83</td>
<td>.96</td>
</tr>
<tr>
<td>Notes</td>
<td>150</td>
<td>.93</td>
<td>.95</td>
</tr>
</tbody>
</table>
# Results of BCL’s Detail Tagging

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Type</th>
<th>Action</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting Policy Note</td>
<td>Full Accounting Note</td>
<td>Detection and Extraction</td>
<td>90%+</td>
</tr>
<tr>
<td></td>
<td>Individual Policy</td>
<td>Detection and Extraction</td>
<td>90%</td>
</tr>
<tr>
<td>Notes Table</td>
<td>Individual Policy</td>
<td>Tagging</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>Individual Table</td>
<td>Detection and Extraction</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td>Table Label</td>
<td>Detection and Extraction</td>
<td>75%</td>
</tr>
<tr>
<td>Detail Notes</td>
<td>Individual Table</td>
<td>Tagging</td>
<td>76%</td>
</tr>
<tr>
<td>Text Notes</td>
<td>Labels</td>
<td>Tagging</td>
<td>70–90%</td>
</tr>
<tr>
<td></td>
<td>Numeric Values</td>
<td>Detection and Extraction</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td>Date</td>
<td>Detection and Extraction</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td>Labels</td>
<td>Detection and Extraction</td>
<td>75%</td>
</tr>
</tbody>
</table>
In this presentation we described identification and extraction of tables, DEI, parenthetical, notes from a 10Q financial document, and overview of detail tagging.

We follow keyword based pattern matching and NLP methods to extract tables, DEI, parentheticals and notes.

For future work:
- We will fine-tune our algorithm to increase the accuracy of parenthetical tables, notes
- Fine-tune NLP methods to detail-tag
- Evaluate XBRL tagging and generation modules
Thank You!

BCL Technologies
3031 Tisch Way, Suite 1000
San Jose, CA 95128, USA
Phone: (408) 557-2080

Demo.smartxbrl.com
www.bclresearch.com