

A Pattern Recognition and NLP Approach to Automated XBRL Extraction

Hassan Alam, Aman Kumar, Cheryl Lee, Yuliya Tarnikova



[BCL Technologies](#)
San Jose, CA 95128



25 Yokohama

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

Company Financial
Source Documents



Data Extraction

SmartXBRL™ 1.1

Source: CEI | Income Statement | Balance Sheet | Balance Sheet Parenthetical | Cash Flow | Cash Flow

Note 3 | Note 6 | Note 7 | Note 8 | Note 9 | Note 10 | Note 11 | Note 12 | Note 13 | Note 14 | Note 15

	2011-07-01	2010-07-01	
INCOME STATEMENTS (Unaudited)			
USD			
in Millions			
Revenue	\$ 17,372	\$ 16,188	us-gaap_Revenue
Operating expenses			us-gaap_OperatingExpenses

Syntactic-Semantic Tagging

```
<context id="Duration_1_1_2010_To_1_1_2010"
  gaap_StatementEquityComponentsAxis="Income Statement"
  >
  <entity
    identifier="1"
    scheme="http://www.sec.gov/CIK"
  >
  <segment
    id="1"
    name="2010"
  >
  <period
    startDate="2010-01-01"
    endDate="2010-12-31"
  >
  </period>
  </segment>
  </entity>
  </context>
```

XBRL Tagged Data

```
gaap:CommonStockMember</xbrl:di
```

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 LABEL LINKBASE	aa-20111231_lab.xml
17	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE	aa-20111231_def.xml
18	XBRL TAXONOMY EXTENSION LABEL LINKBASE	aa-20111231_lab.xml
19	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE	aa-20111231_pre.xml

Government Filing

Data Verification

Flowchart of the SmartXBRL system

Company Financial
Source Documents



Data Extraction



Syntactic-Semantic Tagging

```
<context id="Duration_1_1_2010_To_1_1_2010"
  gaap_StatementEquityComponentsAxis="Net Income"
  >
  <entity
    identifier="123456789"
    scheme="http://www.sec.gov/CIK"
  >
    <segment
      id="1"
      gaap:CommonStockMember</xbrl:
    >
    </segment>
  </entity>
  <period
    startDate="2010-01-01"
    endDate="2010-12-31"
  >
  </period>
</context>
```

XBRL Tagged Data

XBRL Generation

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_ca
17	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE	aa-20111231_def.xml
18	XBRL TAXONOMY EXTENSION LABEL LINKBASE	aa-20111231_lab.xml
19	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE	aa-20111231_pre.xml

Government Filing

Data Verification

Flowchart of the SmartXBRL system

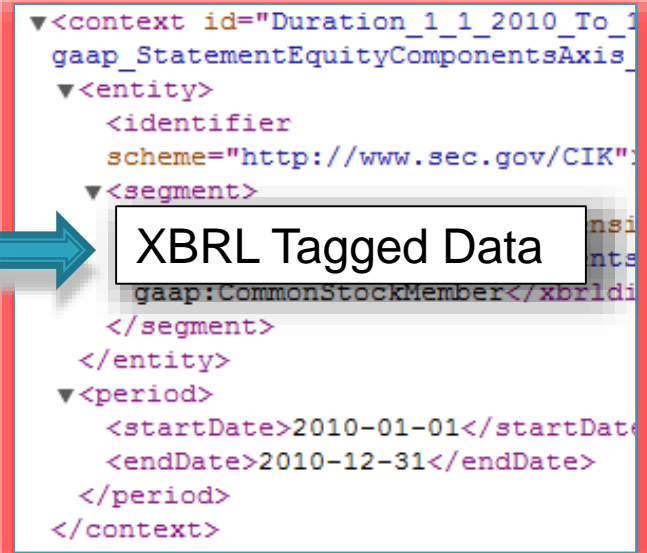
Company Financial
Source Documents



Data Extraction



Syntactic-Semantic Tagging



XBRL Tagged Data

XBRL Generation

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_ca
17	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE	aa-20111231_def.xml
18	XBRL TAXONOMY EXTENSION LABEL LINKBASE	aa-20111231_lab.xml
19	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE	aa-20111231_pre.xml

Government Filing

Data Verification

Flowchart of the SmartXBRL system

Company Financial
Source Documents



Data Extraction

SmartXBRL™ 1.1

	2011-07-01	2010-07-01	
INCOME STATEMENTS (Unaudited)			
USD			
in Millions			
Revenue	\$ 17,372	\$ 16,156	us-gaap_Revenue
Operating expenses			us-gaap_OperatingExpenses

Syntactic-Semantic Tagging

```
<context id="Duration_1_1_2010_To_1_1_2010"
  gaap_StatementEquityComponentsAxis="StatementEquityComponents"
  >
  <entity
    identifier="2010"
    scheme="http://www.sec.gov/CIK"
  >
    <segment
      id="2010-01-01_2010-12-31"
      >
        <period
          startDate="2010-01-01"
          endDate="2010-12-31"
        >
          </period>
        </segment>
      </entity>
    </context>
```

XBRL Tagged Data

```
gaap:CommonStockMember</xbrl:di
```

XBRL Generation

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_ca
17	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE	aa-20111231_def.xml
18	XBRL TAXONOMY EXTENSION LABEL LINKBASE	aa-20111231_lab.xml
19	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE	aa-20111231_pre.xml

Government Filing

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

QUARTERLY FINANCIAL DOCUMENT (10Q)

- ▶ 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.

Sample 10Q

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

FORM 10-Q

(Mark 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

	Third quarter ended		Nine months ended	
	September 30, 2011	September 30, 2010	September 30, 2011	September 30, 2010
Sales (J)	\$6,419	\$5,287	\$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	124
Provision for depreciation, depletion, and amortization	376	358	1,112	1,079
Restructuring and other charges (D)	9	2	49	219
Interest expense	125	139	399	376
Other expenses (income), net (I)	31	43	(47)	48
Total costs and expenses	6,139	5,227	17,660	15,161
Income from continuing operations before income taxes	280	60	1,302	200
Provision (benefit) for income taxes (M)	55	(49)	329	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

(Mark 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

	Third quarter ended		Nine months ended	
	September 30, 2011	2010	2011	2010
Sales (J)	\$6,419	\$5,287	\$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	124
Provision for depreciation, depletion, and amortization	376	358	1,112	1,079
Restructuring and other charges (D)	9	2	49	219
Interest expense	125	139	399	376
Other expenses (income), net (I)	31	43	(47)	48
Total costs and expenses	6,139	5,227	17,660	15,161
Income from continuing operations before income taxes	280	60	1,302	200
Provision (benefit) for income taxes (M)	55	(49)	329	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

(Mark 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

	Third quarter ended		Nine months ended	
	September 30, 2011	September 30, 2010	September 30, 2011	September 30, 2010
Sales (J)	\$6,419	\$5,287	\$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	124
Provision for depreciation, depletion, and amortization	376	358	1,112	1,079
Restructuring and other charges (D)	9	2	49	219
Interest expense	125	139	399	376
Other expenses (income), net (I)	31	43	(47)	48
Total costs and expenses	6,139	5,227	17,660	15,161
Income from continuing operations before income taxes	280	60	1,302	200
Provision (benefit) for income taxes (M)	55	(49)	329	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

(Mark 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

	Third quarter ended		Nine months ended	
	September 30, 2011	September 30, 2010	September 30, 2011	September 30, 2010
Sales (J)	\$6,419	\$5,287	\$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	124
Provision for depreciation, depletion, and amortization	376	358	1,112	1,079
Restructuring and other charges (D)	9	2	49	219
Interest expense	125	139	399	376
Other expenses (income), net (I)	31	43	(47)	48
Total costs and expenses	6,139	5,227	17,660	15,161
Income from continuing operations before income taxes	280	60	1,302	200
Provision (benefit) for income taxes (M)	55	(49)	329	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

DEI Extraction

A sample output of BCL's SmartXBRL DEI extraction

Document and Entity Information (USD \$)	2010-09-01	<Date>	2011-07-15
Document and Entity Information [A]	2011-06-30		
Document Type	10-Q		
Amendment Flag	false		
Document Period End Date	2011-06-30		
Document Fiscal Year Focus	2011		
Document Fiscal Period Focus	Q4		
Entity Registrant Name			

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*>

Performance – DEI Extraction

- ▶ 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

Entity Public Float (for 10-K)				dei_EntityPublicFloat dei_EntityPublicFloat State aggregate market value computed by reference to price
Entity Common Stock, Shares Outstanding (for 10-K, 10-Q, 20-F, 40-F)				dei_EntityCommonStock dei_EntityCommonStockShare Indicate number of shares out of latest practicable date. Y/N

2010-01-01	<Date>	<Date>	=
2010-03-31			
10-Q			dei_DocumentType dei_DocumentType The type of document being limited to the same value as I
false			dei_AmendmentFlag dei_AmendmentFlag If the value is true, then the

FINANCIAL TABLE EXTRACTION

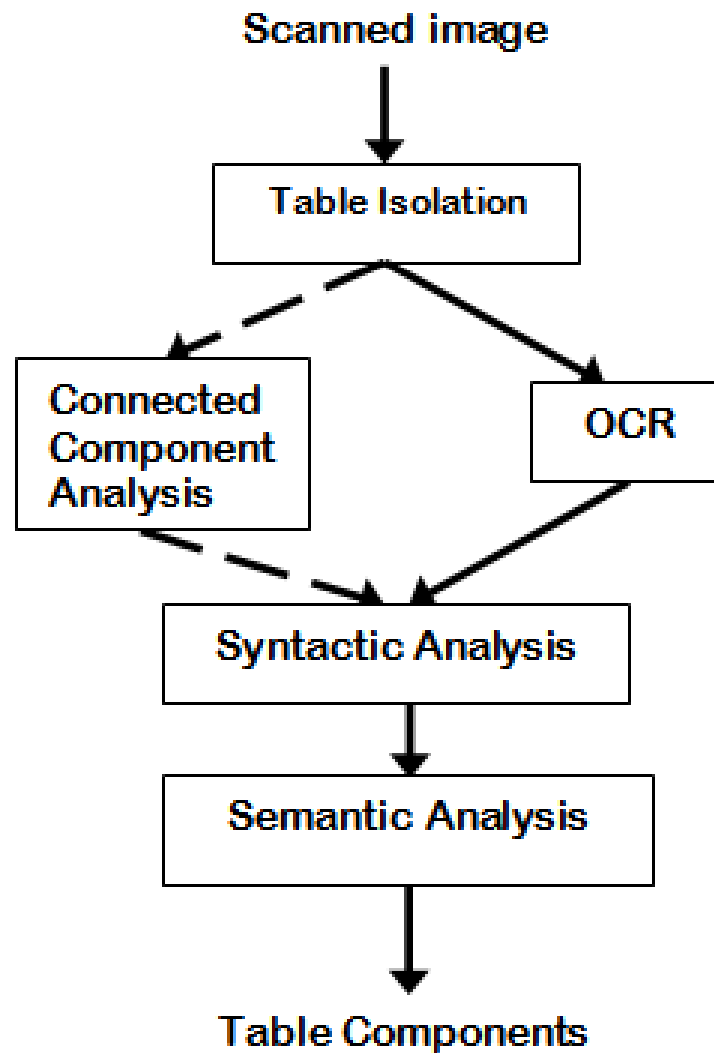
Tasks

- ▶ 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.)

Previous Work (Tupaj et. al., 1996)

- ▶ 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 Contd.



Previous Work Contd.

In this work we are able to:

- ▶ Isolate embedded tables from documents
- ▶ Identify table components such as title blocks, table entries, and footer blocks

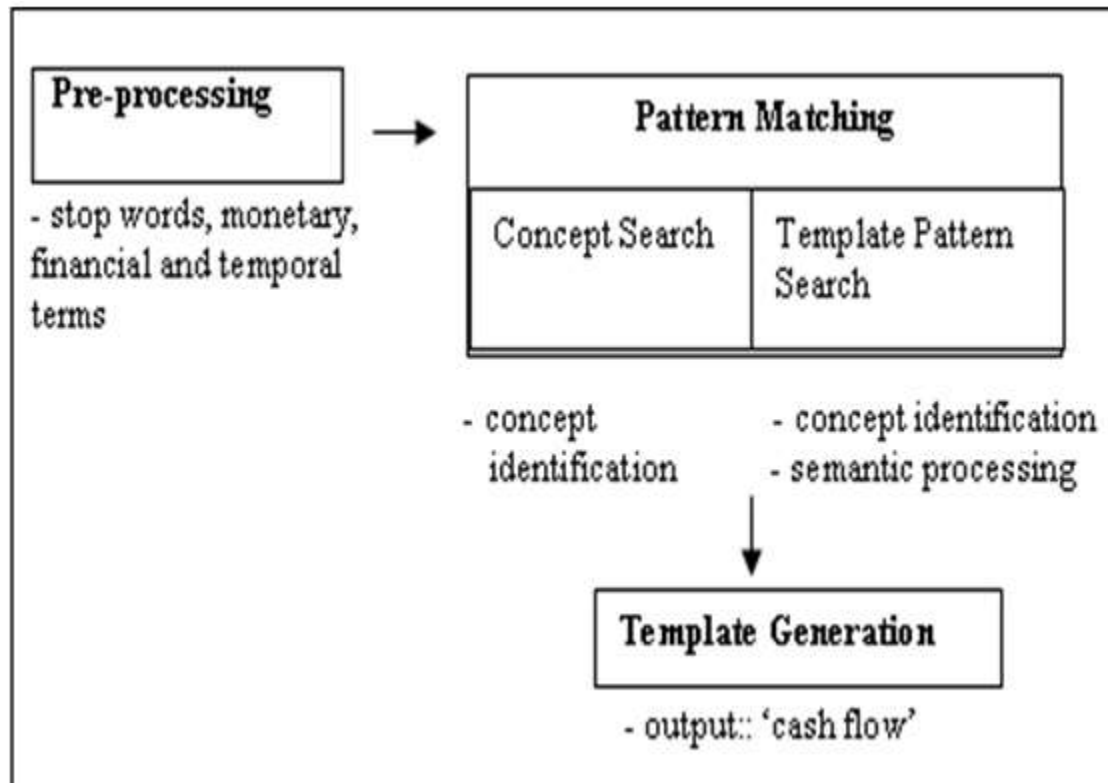
Performance – Previous work

	Number of Samples	Detection of Tables		Detection of Cells	
		P	R	P	R
Scientific Journals	100	0.930	0.924	0.970	0.916
Financial Tables	100	0.714	0.750	0.920	0.890

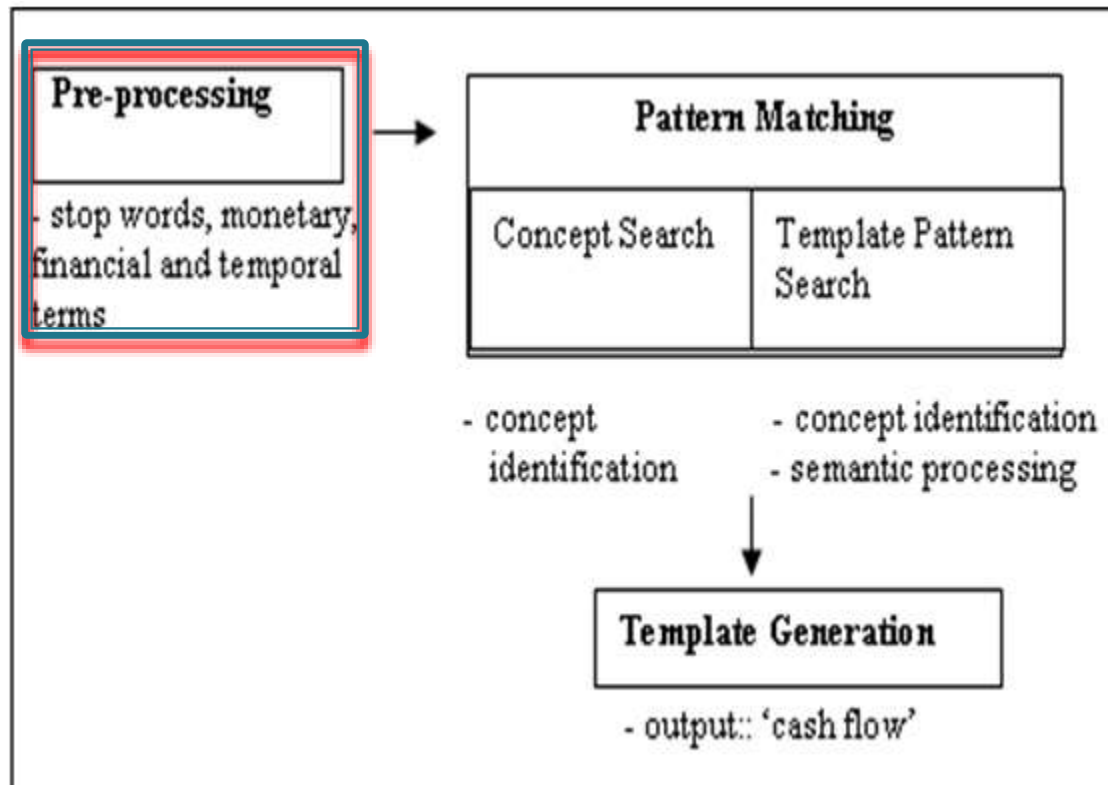
Financial Table Extraction – Current System

- ▶ 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.

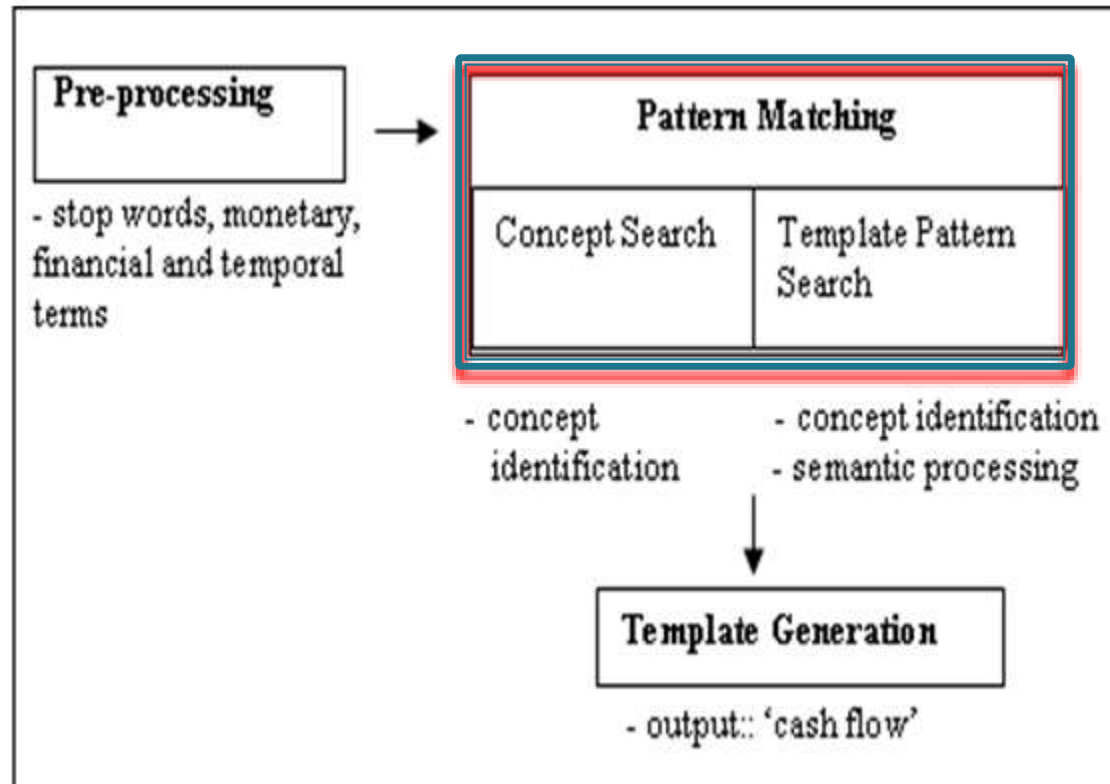
Extracting Four Face Financial Tables



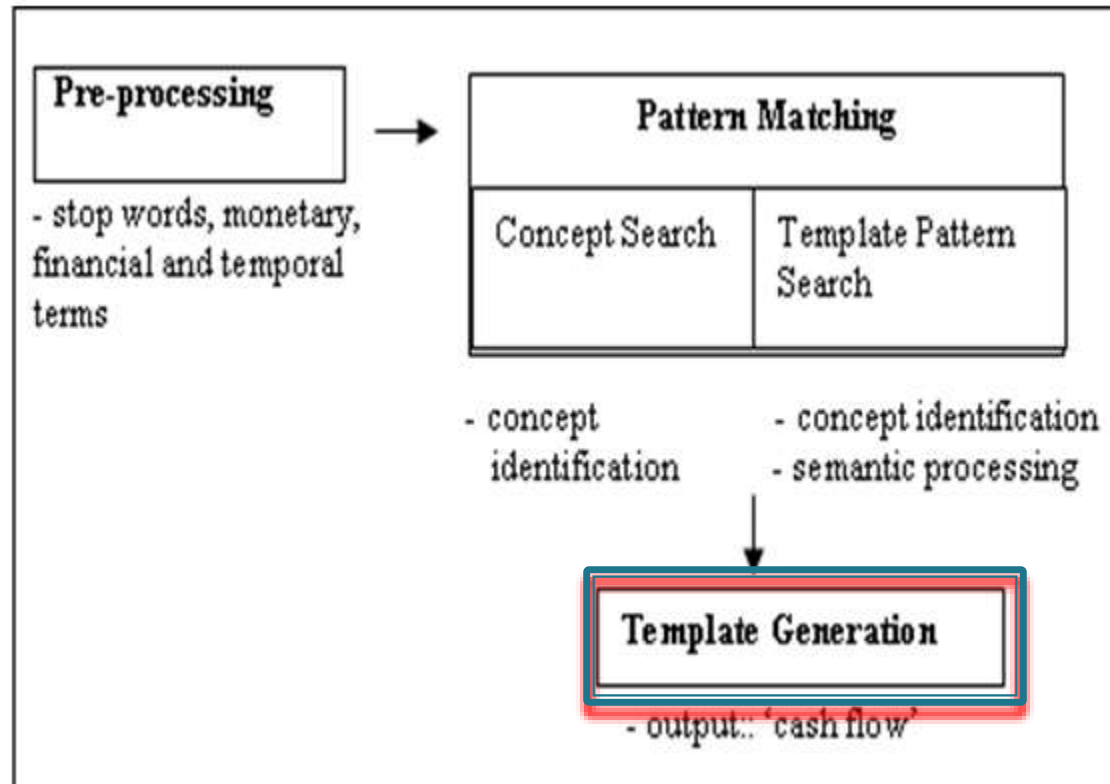
Extracting Four Face Financial Tables



Extracting Four Face Financial Tables



Extracting Four Face Financial Tables



GENERATE AUTO PARENTHETICALS

- ▶ 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

(in thousands)	2007	2006
Receivables, less allowances of \$1,260 and \$1,150	31,659	31,601



- String matching algorithm
- Natural Language Processing methods

(in thousands)	2007	2006
Receivables, less allowances	31,659	31,601
(Less allowances)	(1,260)	(1,150)

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

SmartXBRL™ 1.0

[Document](#) [Filing Properties](#) [Payment](#) [Sign](#)

Source	DEI	Income Statement	Balance Sheet	Cash Flow	Equity	Equity Parenthetical	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

[\[Add Abstract Row\]](#)

	2010-09-01 ?
Basis of Presentation - The interim	
	2011-06-30 ?

Basis of Presentation - The interim Cons	<!--DOCTYPE html PUBLIC "-//W3C/DTD XHTML 1.0 Transitional/EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd"><html><head><meta charset="utf-8"><title></title></head><body><div style="border: 1px solid black; padding: 5px;"><table border="1" style="width: 100%; border-collapse: collapse;"><tr style="background-color: #e6f2ff;"><td style="width: 30%; padding: 5px;">Basis of Presentation - The interim</td><td style="padding: 5px;"><div style="font-size: 10pt;">A. Basis of Presentation The interim Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (Alcoa) 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.</div></td></tr></table></div></body></html>
--	---

Snapshot of Notes output by SmartXBRL

SmartXBRL™ 1.0

[Document](#) [Filing](#) [Properties](#) [Payment](#) [Sign](#)

Source	DEI	Income Statement	Balance Sheet	Cash Flow	Equity	Equity Parenthetical	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

[\[Add Abstract Row\]](#)

Basis of Presentation - The interim	2010-09-01 ?
	2011-06-30 ?

Basis of Presentation - The interim Cons: <!--DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
A. Basis of Presentation
The interim Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (Alcoa; Alcoa Company; 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.

Snapshot of Notes output by SmartXBRL

SmartXBRL™ 1.0

[Document](#) [Filing Properties](#) [Payment](#) [Sign](#)

Source	DEI	Income Statement	Balance Sheet	Cash Flow	Equity	Equity Parenthetical	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

[\[Add Abstract Row\]](#)

	2010-09-01 ?
Basis of Presentation - The interim	
	2011-06-30 ?

Basis of Presentation - The interim Cons: <!--DOCTYPE html PUBLIC "-//W3C/DTD XHTML 1.0 Transitional/EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd"><html><head><meta charset="utf-8"><title></title></head><body><div style="border: 1px solid black; padding: 5px;"><table border="1" style="width: 100%; border-collapse: collapse;"><tr><td style="width: 30%; padding: 5px;">Basis of Presentation - The interim</td><td style="padding: 5px;"><p>Basis of Presentation - The interim</p><p>Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (Alcoa) 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.</p></td></tr></table></div></body></html>

Snapshot of Notes output by SmartXBRL

SmartXBRL™ 1.0

[Document](#) [Filing Properties](#) [Payment](#) [Sign](#)

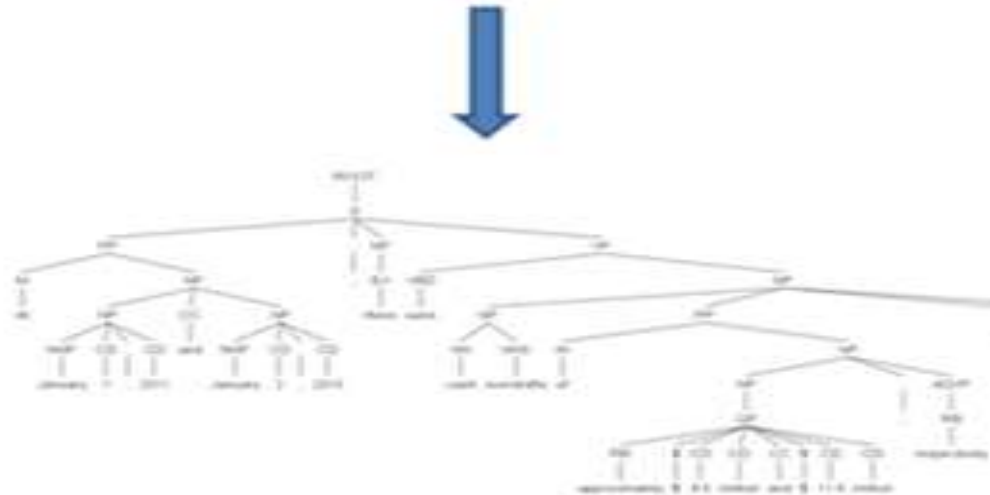
Source	DEI	Income Statement	Balance Sheet	Cash Flow	Equity	Equity Parenthetical	Note 1	Note 2	Note 3	Note 4		
Note 5	Note 6	Note 7	Note 8	Note 9	Note 10	Note 11	Note 12	Note 13	Note 14	Note 15	Note 16	Note 17

[\[Add Abstract Row\]](#)

	2010-09-01 ?
Basis of Presentation - The interim	
	2011-06-30 ?
Basis of Presentation - The interim Cons	<!--DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd"><html><head><meta charset="utf-8"><title></title></head><body><div style="border: 1px solid black; padding: 5px;"><table border="1" style="width: 100%; border-collapse: collapse;"><tr style="background-color: #e6f2ff;"><td style="width: 30%; padding: 5px;">Basis of Presentation - The interim</td><td style="padding: 5px;"><div style="text-align: right; font-weight: bold; font-size: 10pt; margin-bottom: 5px;">2010-09-01 ?</div><div style="text-align: right; font-weight: bold; font-size: 10pt; margin-bottom: 5px;">2011-06-30 ?</div><table border="1" style="width: 100%; border-collapse: collapse; margin-top: 5px;"><tr style="background-color: #e6f2ff;"><td style="width: 30%; padding: 5px;">Basis of Presentation - The interim</td><td style="padding: 5px;"><div style="font-size: 10pt; padding: 5px;"><p>Basis of Presentation - The interim</p><p>Consolidated Financial Statements of Alcoa Inc. and its subsidiaries (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.</p></div></td></tr></table></div></body></html>

Detail Note Tagging – Overall schema

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



Consolidated Balance Sheets (USD \$) In Millions	Jan. 01, 2011	Jan. 02, 2010
Accounts Payable		
Cash Overdraft	\$ 9.0	\$ 11.6

RESULTS of XBRL Data Extraction

	Number of 10-Q Samples	Detection		Extraction of Cells/Content	
		P	R	P	R
Tables	150	.96	.95	.96	.97
DEI	150	.97	.95	.93	.96
Parenthetical	150	.83	.96	.78	.87
Notes	150	.93	.95	.86	.95

Results of BCL's Detail Tagging

Item	Item Type	Action	Accuracy
Accounting Policy Note	Full Accounting Note	Detection and Extraction	90%+
	Individual Policy	Detection and Extraction	90%
Notes Table	Individual Policy	Tagging	75%
	Individual Table	Detection and Extraction	95%
	Table Label	Detection and Extraction	75%
Detail Notes	Individual Table	Tagging	76%
	Labels	Tagging	70-90%
Text Notes	Numeric Values	Detection and Extraction	95%
	Date	Detection and Extraction	95%
	Labels	Detection and Extraction	75%

CONCLUSIONS AND FUTURE WORK

- ▶ 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](#)

