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



		Example t id="Duration 1 1 2010 To
	SmartXBRL 1.1	gaap_StatementEquityComponentsAxis
str Market	Der Source (DEL Drome Dietement Belance Sheet Belance Sheet Parenthetical Caol Flow Caol Flow Note 5 Note 6 Note 7 Note 8 Note 5 Note 10 Note 11 Note 12 Note 12 Note 13 Note 14 Note 13	▼ <entity> <identifier scheme="http://www.sec.gov/CIK"</identifier </entity>
	+0+ +0+	<pre>v<segment></segment></pre>
	INCOME STATEMENTS (Invariantee) 2011 array 0 2016 array 0 - 0	XBRL Tagged Data
	Annual State	<pre>gaap:CommonStockMember </pre>
1.1773.01	R «Custoriza»	* Cherions
HIML	Operating expension Diff Diff Diff Diff Diff Diff Diff Diff Diff Di	<pre><startdate>2010-01-012010-12-31 </startdate></pre>
	XBRL Generation	<pre><startdate>2010-01-012010-12-31 </startdate></pre>
	Approximation of the second se	<pre><startdate>2010-01-012010-12-31 </startdate></pre>
	Operating expenses Interview Construction Interview Construction Seq Description Description Document 14 XBRI INSTANCE DOCUMENT Seq-20111231 xml	<pre><startdate>2010-01-01<enddate>2010-12-31</enddate> Data Verification</startdate></pre>
	Operating expenses Last See See See See See See See See See Se	<pre><startdate>2010-01-01<enddate>2010-12-31</enddate> Data Verification</startdate></pre>
	Determine the present of the presen	<pre><startdate>2010-01-012010-12-31 Data Verification </startdate></pre>
	See Description Document 14 XBRL INSTANCE DOCUMENT aa-20111231_def.xml 15 XBRL TAXONOMY EXTENSION DEFINITION aa-20111231_def.xml	<pre><startdate>2010-01-01<enddate>2010-12-31</enddate> Data Verification</startdate></pre>
	Operating expenses Int Int </td <td><pre><startdate>2010-01-01<enddate>2010-12-31</enddate> Data Verification</startdate></pre></td>	<pre><startdate>2010-01-01<enddate>2010-12-31</enddate> Data Verification</startdate></pre>



SmartXBRL"1.1 Image: StatementEquityComponentsAxis Image: StatementEquityComponentsAxis <th></th> <th></th> <th></th>			
Status		SmartXBRL 1.1	<pre>v<context cik")<="" http:="" id="Duration_1_1_2010_To_1 gaap_StatementEquityComponentsAxis_</pre></th></tr><tr><td>Segments Segments Segments</td><td>st 🚺 💭</td><td>Do
Source (DEL Territe Distance Direct Balance Direct Territatical Call Flow (Call Flow
Note 3 Note 4 Note 5 Note 5 Note 10 Note 12 Note 12 Note 12 Note 13 Note 14 Note 13 </td><td><pre>▼<entity> <identifier scheme=" pre="" www.sec.gov=""></context></pre>
Subscription Data Files Seq Description Document 14 XBRL INSTANCE DOCUMENT as-20111231.xml 15 XBR TAXONIONY EXTENSION SOLIEMA as-20111231.xml 16 XBR Government Filing a1_ca 17 XBRL TAXONOMY EXTENSION DEFINITION as-20111231_def.xml 18 XBRL TAXONOMY EXTENSION LABEL LINKBASE as-20111231_def.xml 19 XBRL TAXONOMY EXTENSION PRESENTATION as-20111231_gab.xml 19 XBRL TAXONOMY EXTENSION PRESENTATION as-20111231_gab.xml	Source Docume	Image: Statement Image: Statement Image: Statement Image: Statement	<pre></pre>
Seq Description Document 14 XBRL INSTANCE DOCUMENT aa-20111231.xml 15 XBRL COVERNMENT EXTENSION EXCLEMA co-20111231.xsd 16 XBRL GOVERNMENT FILING 31_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 aa-20111231_reg xml			
		XBRL Generation	



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.



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 September 30,		Nine mon Septem	ths ended iber 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 Fom 10-Q report should be read in conjunction with Alcoa's Annual Report on Fom 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.



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 mon	ths ended
	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 Fom 10-Q report should be read in conjunction with Alcoa's Annual Report on Fom 10-K for the year ended December 31, 2010, which includes all disclosures required by GAAP.



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 en September 30. September 3		ths ended ber 30,	
	2011	2010	2011	2010
Sales (J)	\$6,419	\$5,287	\$18,962	ne months ended Spetember 30, 2010 2010 962 \$15,361 12,552 12,636 759 679 136 124 112 1,079 49 219 399 376 (47) 48 (660 15,161 302 200
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.



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 September 30,		Nine mon Septem	ths ended ber 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.



DEI Extraction

A sample output of BCL's SmartXBRL DEI extraction

Document and Entity Information	2010-09-01 🕚	<date></date>	2011-07-15
(USD \$)	2011-06-30 🕚		
Document and Entity Information (A			
Document Type	10-0		
Artendment Flag	tabe		
Document Period End Date	2011-06-30		
Document Fiscel Year Focus	2011		
Document Fiscal Period Focus	Q4		
Ertity 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

		2010-01-01 ()	<date></date>	«Date»	=@
Entry Public Float	dei_ErstyPublicFloat	2010-01.31			
(Bur 10-8)	der ErstyPusichlost. State approprie market volue computed by reference to pric	10-9			dei_DocumentType
Entity Common Stock, Shares Outstandin	der_EntityCommonStock				del_DocumentType The type of document being limited to the same value as
(bu 184, 18-Q, 28-F, 48-F)	dei_EnthyConnorStockShare Indicate number of shares out	Twise			dei_AmendmentFlag
	of intest practicable date. Why				del_AmendmentFlag. If the value is true, then the r



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	
		Ρ	R	Р	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.



















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 S	tatement	Balanci	e Sheet	Cash Flow	Equity	Equity Pa	renthetical	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 1	17
(<u>Add</u> Basis Basis	Abstra s of Pres	act Row] esentation -	The interi	m Ist dDO margin Preset Conso (8#82 Conso recurr	CTYPE htm i-left:0ir(ma ildated Fina 20;Alcoa&d ildated Fina ildated Fina	I PUBLIC *-/// argin-right:0in; 211; - incial Stateme #8221; or the incial Stateme ents, conside	/3C//DTD X8 margin-top: v/b>The inte v/b>The inte v/b}The inte <tr< th=""><th>HTML 1.0 Tra 9.0pt,"> rim a Inc. and its nparry8#8222 all adjustmer ary by mana</th><th>nstional//EN* span style="1 subsidiaries 1;) are unauc nts, consistin gement to fa</th><th>201 201 "http://www font-size:10 itted, These g of normal inly state the</th><th>0-09-01 (2 1-06-30 (2 w.w3.org/T I.Opt,">A. B:</th><th>) R/xhtml1/L asis of</th><th>)TD/xhtm</th></tr<>	HTML 1.0 Tra 9.0pt,"> rim a Inc. and its nparry8#8222 all adjustmer ary by mana	nstional//EN* span style="1 subsidiaries 1;) are unauc nts, consistin gement to fa	201 201 "http://www font-size:10 itted, These g of normal inly state the	0-09-01 (2 1-06-30 (2 w.w3.org/T I.Opt,">A. B:) R/xhtml1/L asis of)TD/xhtm
				Compa result: indical year-e does r accep be rea year e GAAP	any’, s reported i tive of the r and balance hot include ted in the U id in conjun inded Dece	s results of o n these Consi esults that ma s sheet data v all disclosure: Inited States o ction with Alo mber8#160,3 p>	perations, olidated Fini ay be expec- vas derived s required b of America (coa8#8217; 1, 2010, wh	financial post ancial Statem ted for the e I from audited y accounting (GAAP). This s Annual Rep sich includes	tion, and cas ents are not ntire year. Tr d financial sta g principles g Form 10-Q r port on Form all disclosure	h flows: Th necessarily te 2010 tements bu enerally eport shoul 10-K for the es required	e t d by		



SmartXBRL 1.0

									Document	Filing Pr	operties	Payment	t <u>Sign</u>
Source	DEI	Income S	tatement	Balance	e Sheet	Cash Flow	Equity	Equity Pa	renthetical	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 1	7
[Add	Abstra	ect Rowl											
					2010-09-01 (2)								
Basi	Basis of Presentation - The interim			n	2011-06-30 🙁								
Basis of Presentation - The interim Const				Ist et-DO margin Preser Conso (8#82; Conso recurr Compe results indicat year-e does r accept be rea year e GAAP	CTYPE htm Heft: Oin; ma- itation 8#8 idated Fina- ing adjustm any8#8217; a reported i ive of the r and balance tot include ted in the U d in conjun nded Dece 	I PUBLIC */// ingin-right:0in; 211; incial Stateme (8221; or the l incial Stateme ents, conside (s results of o n these Consi- esuits that me is sheet data v all disclosure: hited States o ction with Alo mber 3 p>	ASC/IDTD X8 margin-top: Ab>The intents of Alco 8#8220;Corr offs include red necess perations, 1 olidated Finu ay be expect vas derived s required b of America (oa8#8217; 1, 2010, wf	HIML 1.0 Tra 9.0pt;"> 9.0pt;">9.0pt;"><td< th=""><td>nstional/EW span style=" subsidiaries 1;) are unauc its, consistin gement to fa tion, and cas ents are not ntire year. Tr financial sta principles g Form 10-Q r port on Form all disclosure</td><th>itted. These g of normal inly state the h flows. Th necessarily he 2010 atements bu enerally report shoul 10-K for the es required</th><th>w.w3.org/1 L0pt,">A.Ba e e t d by</th><th>Roxitemita isis of</th><td>TD/schtm</td></td<>	nstional/EW span style=" subsidiaries 1;) are unauc its, consistin gement to fa tion, and cas ents are not ntire year. Tr financial sta principles g Form 10-Q r port on Form all disclosure	itted. These g of normal inly state the h flows. Th necessarily he 2010 atements bu enerally report shoul 10-K for the es required	w.w3.org/1 L0pt,">A.Ba e e t d by	Roxitemita isis of	TD/schtm



Snapshot of Notes output by SmartXBRL SmartXBRL 1.0

Document Filing Droparties Doumant Sign

								4	Document	r ming r n	openies	aymer	it oldin
Source	DEI	Income S	Statement	Balanc	e Sheet	Cash Flow	Equity	Equity Pa	renthetical	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 1	.7
[Add	Ahetra	et Rowl											70
1.100	10000			201						201	0-09-01 (2	(
Basi	s of Pre	sentation	- The interi	m						201	1-06-30 🙋	(
clasis of Presentation - The Internet Const				margii Prese Const (8#82 Const recun Comp result indica year-t does accep be rea year o GAAF	n-left: Oin, ma Intation : Nidated Fina 20; Alcoa&# Nidated Fina ing adjustm any' s reported in tive of the n end balance not include i ted in the U id in conjun ended Dece <to><to><to><to><to><to><to><to><to><to></to></to></to></to></to></to></to></to></to></to>	rgin-right:0in, 211; «/span»- ncial Stateme 8221; or the i ncial Stateme ents, conside s results of o n these Consi esuits that me sheet data v all disclosure: nited States o ction with Alo mber 3	margin-top: «/b>The inte ints of Alco "Cor ints include irred necess operations, 1 olidated Finit ay be expect vas derived s required b of America (coa8#8217; 1, 2010, wf	9.0pt,"> erim a Inc. and its inpany8#8222 all adjustmer iary by mana financial position audited from audited ted for the electron for audited y accounting (GAAP). This is Annual Rep hich includes 	span style="1 subsidiaries 1;) are unaud nts, consistin igement to fai tion, and cas ients are not ntire year. Th d financial sta g principles gr Form 10-Q r port on Form all disclosure	ited. These g of normal rly state the h flows. Th necessarily e 2010 tements bu enerally eport shoul 10-K for the is required	l0pt,">A.B: e e t by	isis of	



SmartXBRL 1.0

Decument Filing Dranartice Dowmant Cign

ource DEI Income Statement E ote 5 Note 6 Note 7 Note 8 N	Valance Sheet Cash Flow Equity Equity Parenthetical ote 9 Note 10 Note 11 Note 12 Note 13 Note 14	Note 1 Note 15	Note 2 Note 16	Note 3 Note 1	Note
[Add Abstract Row]					201
Dania of Dragonstation. The interim		2010	0-09-01 📀		
basis of Presentation - The Interim		2011	1-06-30 🕐		
	margin-left.0in(margin-right:0in(margin-top:9.0pt;"> <span "company")="" alcoa"="" are="" or="" style="f
Presentation – The interim
Consolidated Financial Statements of Alcoa Inc. and its subsidiaries
(" the="" unaud<br="">Consolidated Financial Statements include all adjustments, consisting recurring adjustments, considered necessary by management to fair Company's results of operations, financial position, and cash results reported in these Consolidated Financial Statements are not r indicative of the results that may be expected for the entire year. Th year-end balance sheet data was derived from audited financial stat does not include all disclosures required by accounting principles ge accepted in the United States of America (GAAP). This Form 10-Q re be read in conjunction with Alcoa's Annual Report on Form 10 year ended December 31, 2010, which includes all disclosure	ont-size:10 ted. These of normal rly state the h flows. The necessarily e 2010 tements but enerally eport should IO-K for the is required t	Opt,">A. Ba	sis of	



Detail Note Tagging - Overall schema

At January 1, 2011 and January 2, 2010, there were cash overdrafts of approximately 59.0 million and \$11.6 million, respectively, which are included in accounts payable within the accompanying consolidated balance sheets





RESULTS of XBRL Data Extraction

	Number of 10-Q Samples	Detection		Extraction of Cells/Content		
		Р	R	Ρ	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

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



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

