Improving Data Quality for Web Services Composition

Xitong Li, Stuart Madnick, Hongwei Zhu, Yushun Fan

Stuart E. Madnick

MIT Sloan School of Management

smadnick@mit.edu

The 7th International Workshop on Quality in Databases (QDB'09)
August 24th, 2009, Lyon, France

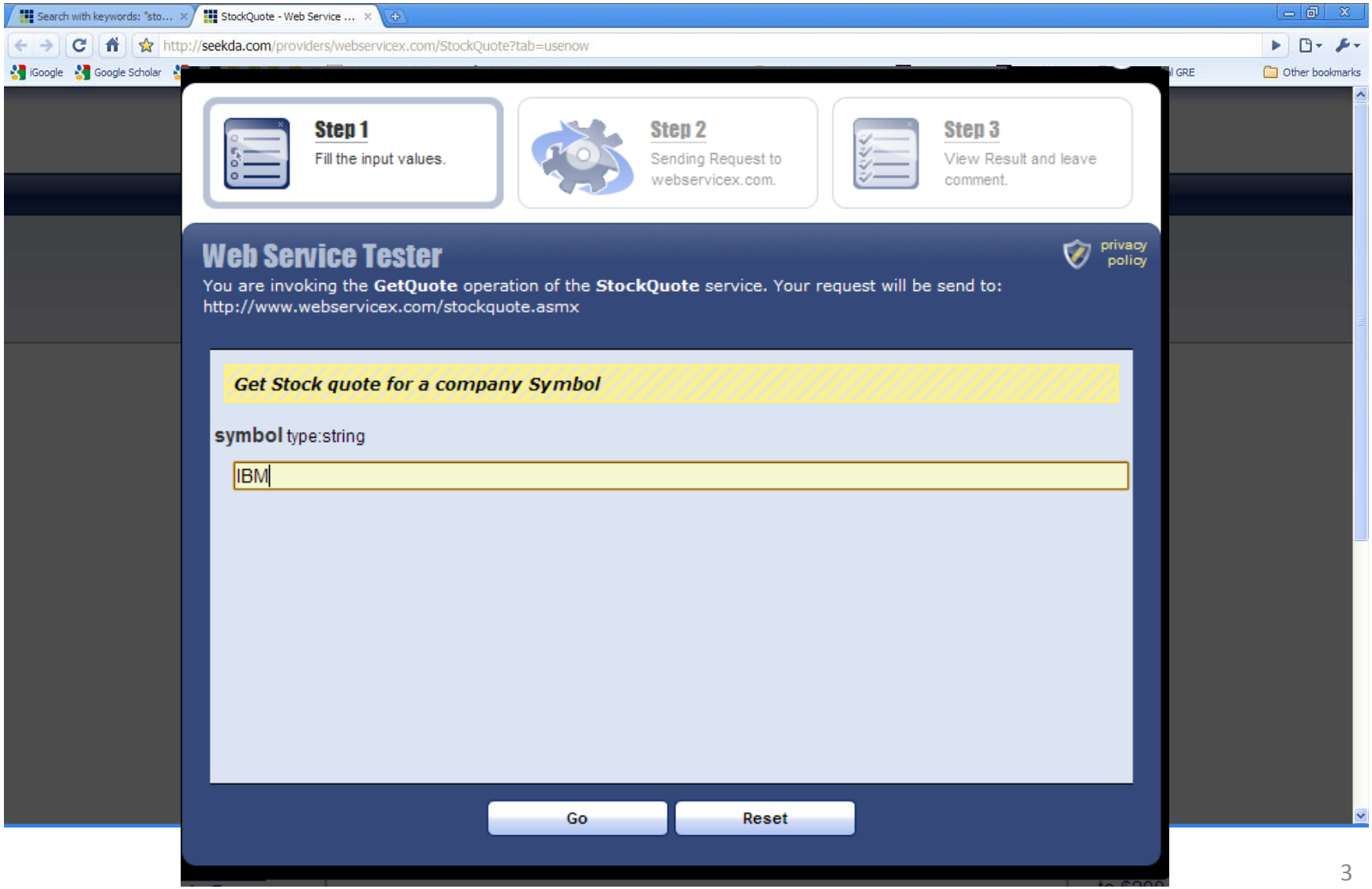
Agenda

- **Research Motivation** – *focused on data quality*
  - Problem of **data misinterpretation:**

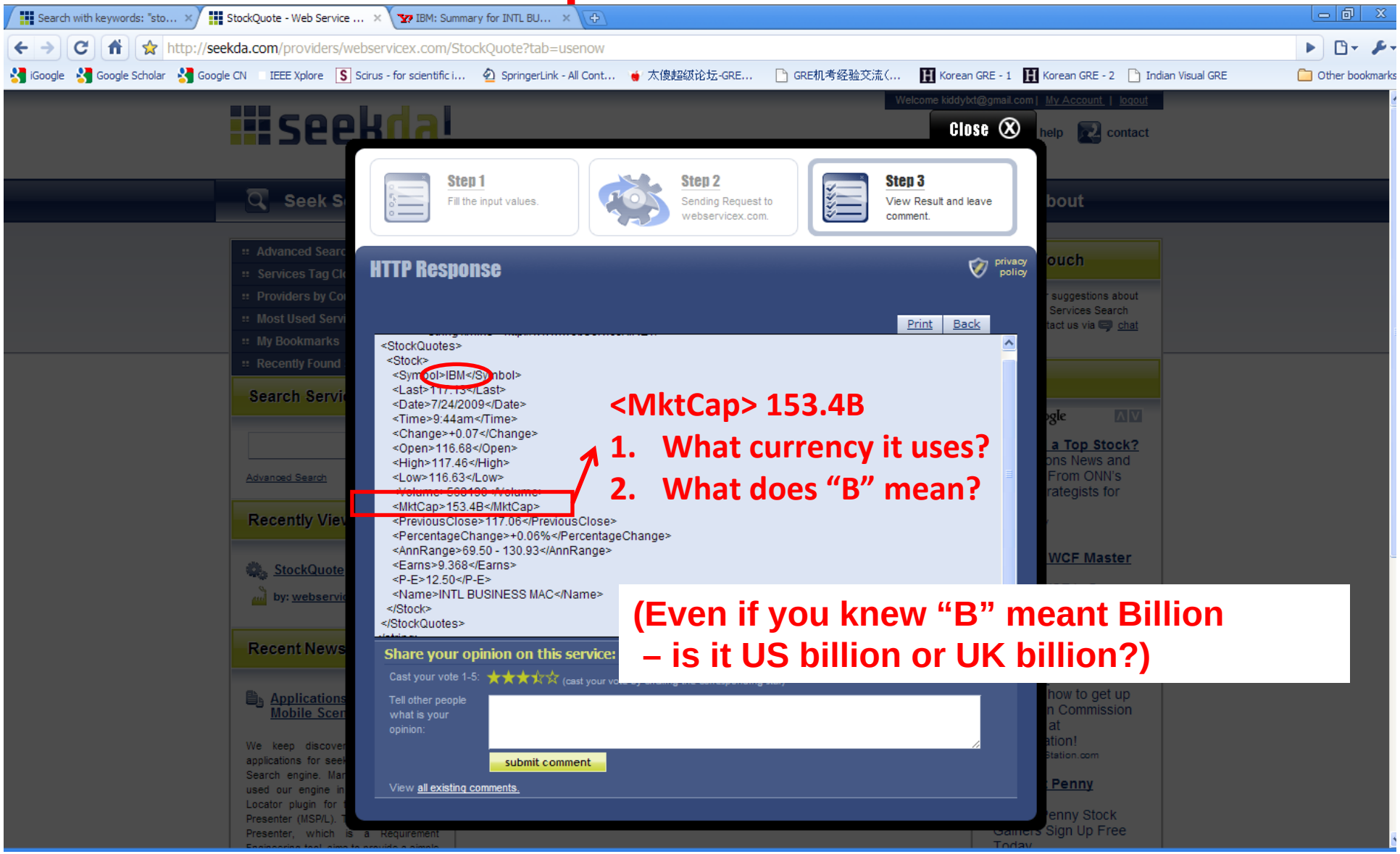
    Different Web services have different assumptions about the interpretation of the exchanged data among them in the composition
- **Our Solution Approach**
  - Representation of Ontology and Context
  - Semantic Annotation
  - Reconciliation Approach
  - Prototype: *CMT*
  - Summary and Future Work

Real-world Scenarios:

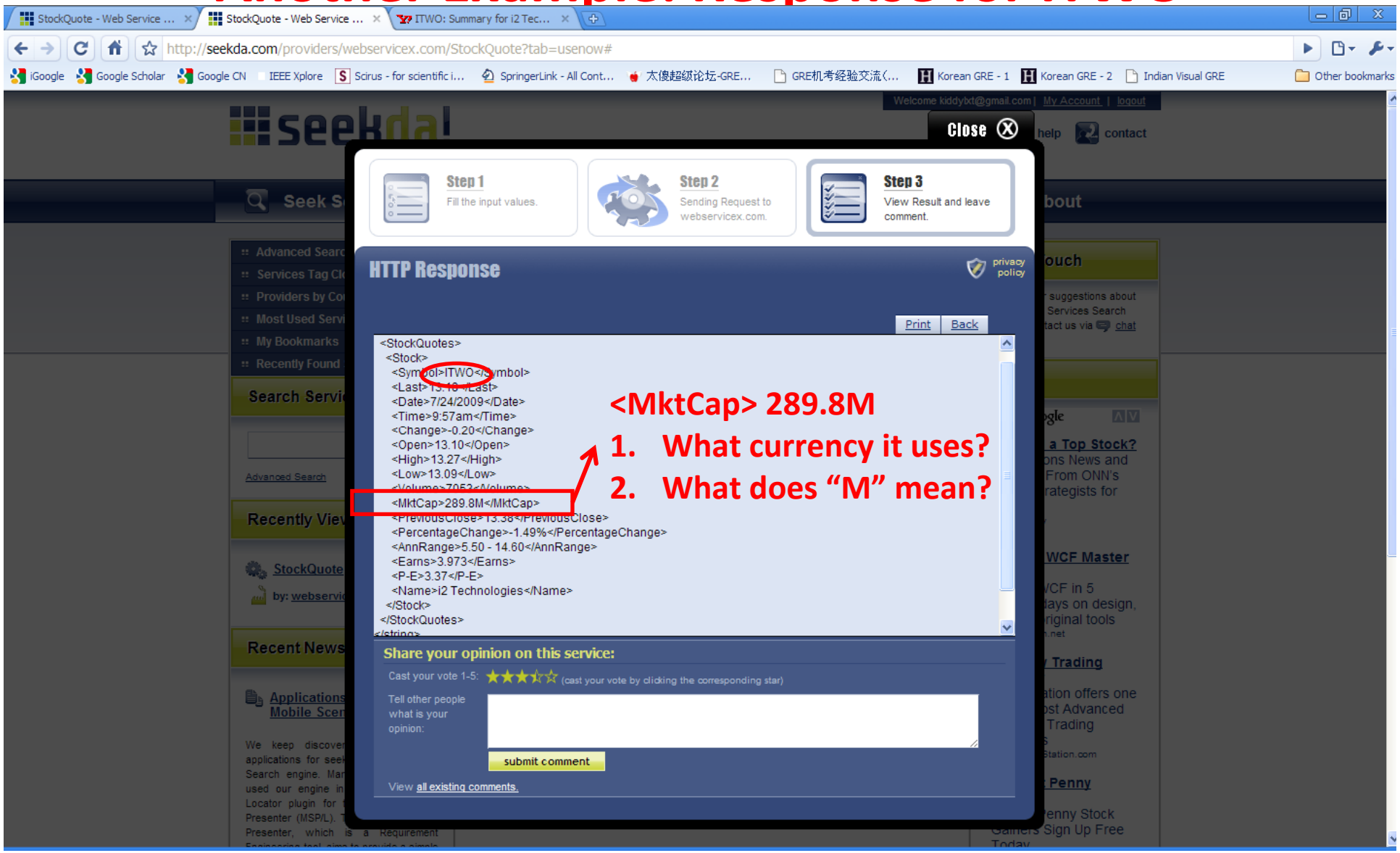
GetQuote operation of the StockQuote service



Response for IBM

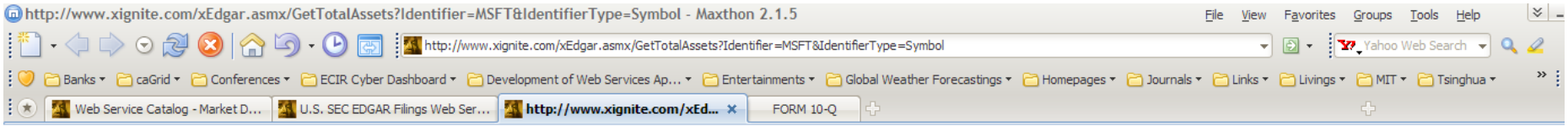


Another Example: Response for ITWO



Different Example: Xignite Web Service

GetTotalAssets for MSFT

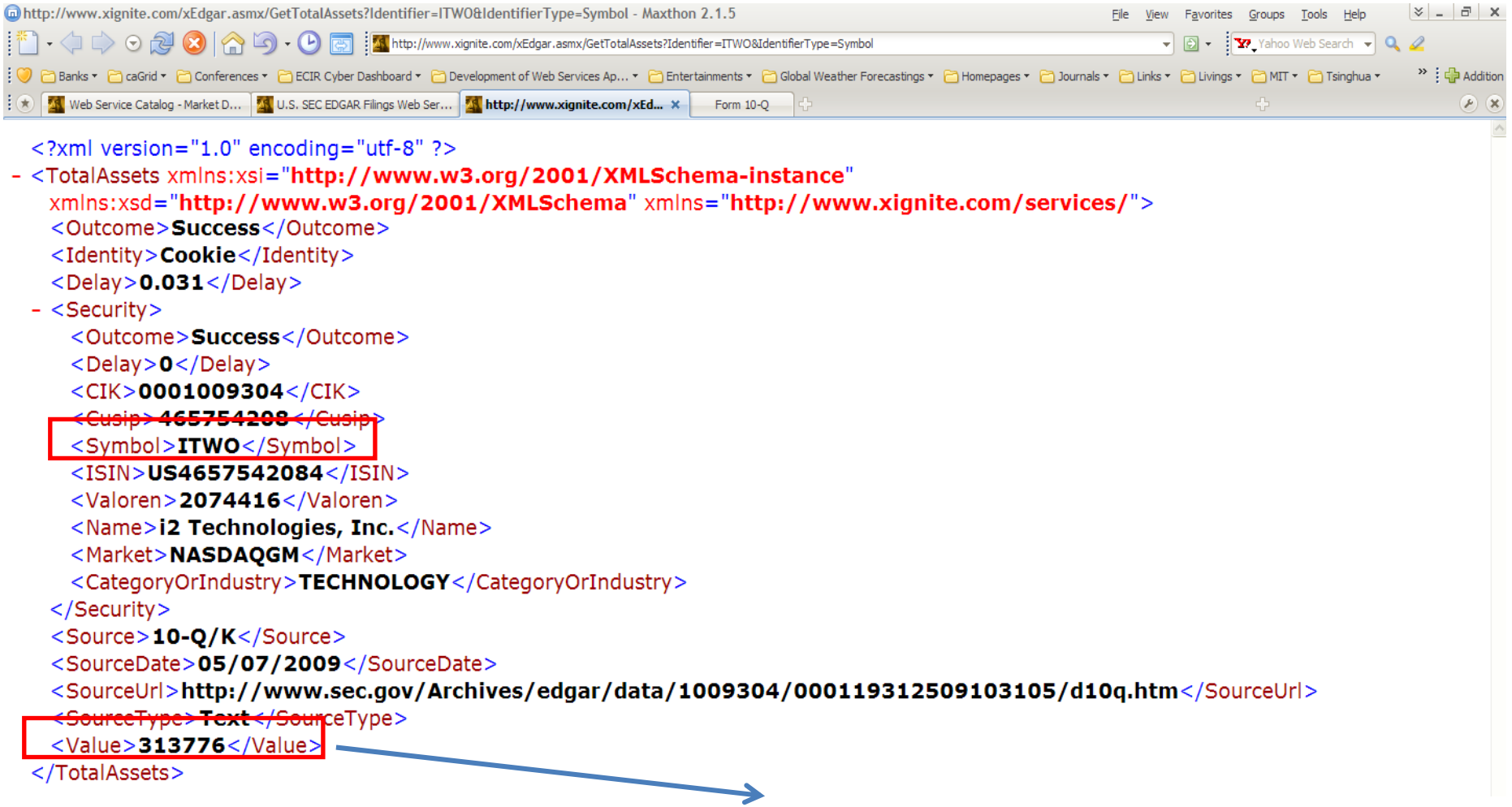



```
<?xml version="1.0" encoding="utf-8" ?>
- <TotalAssets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns="http://www.xignite.com/services/">
    <Outcome>Success</Outcome>
    <Identity>Cookie</Identity>
    <Delay>0.016</Delay>
  - <Security>
      <Outcome>Success</Outcome>
      <Delay>0</Delay>
      <CIK>0000789019</CIK>
      <Cusip>594918104</Cusip>
      <Symbol>MSFT</Symbol>
      <ISIN>US5949181045</ISIN>
      <Valoren>951692</Valoren>
      <Name>Microsoft Corporation</Name>
      <Market>NASDAQGS</Market>
      <CategoryOrIndustry>TECHNOLOGY</CategoryOrIndustry>
    </Security>
    <Source>10-Q/K</Source>
    <SourceDate>04/23/2009</SourceDate>
    <SourceUrl>http://www.sec.gov/Archives/edgar/data/789019/000119312509085779/d10q.htm</SourceUrl>
    <SourceType>Text</SourceType>
    <Value>68853</Value>
  </TotalAssets>
```

What is this date 04/23/2009 ?
What if it was 04/05/06 ?

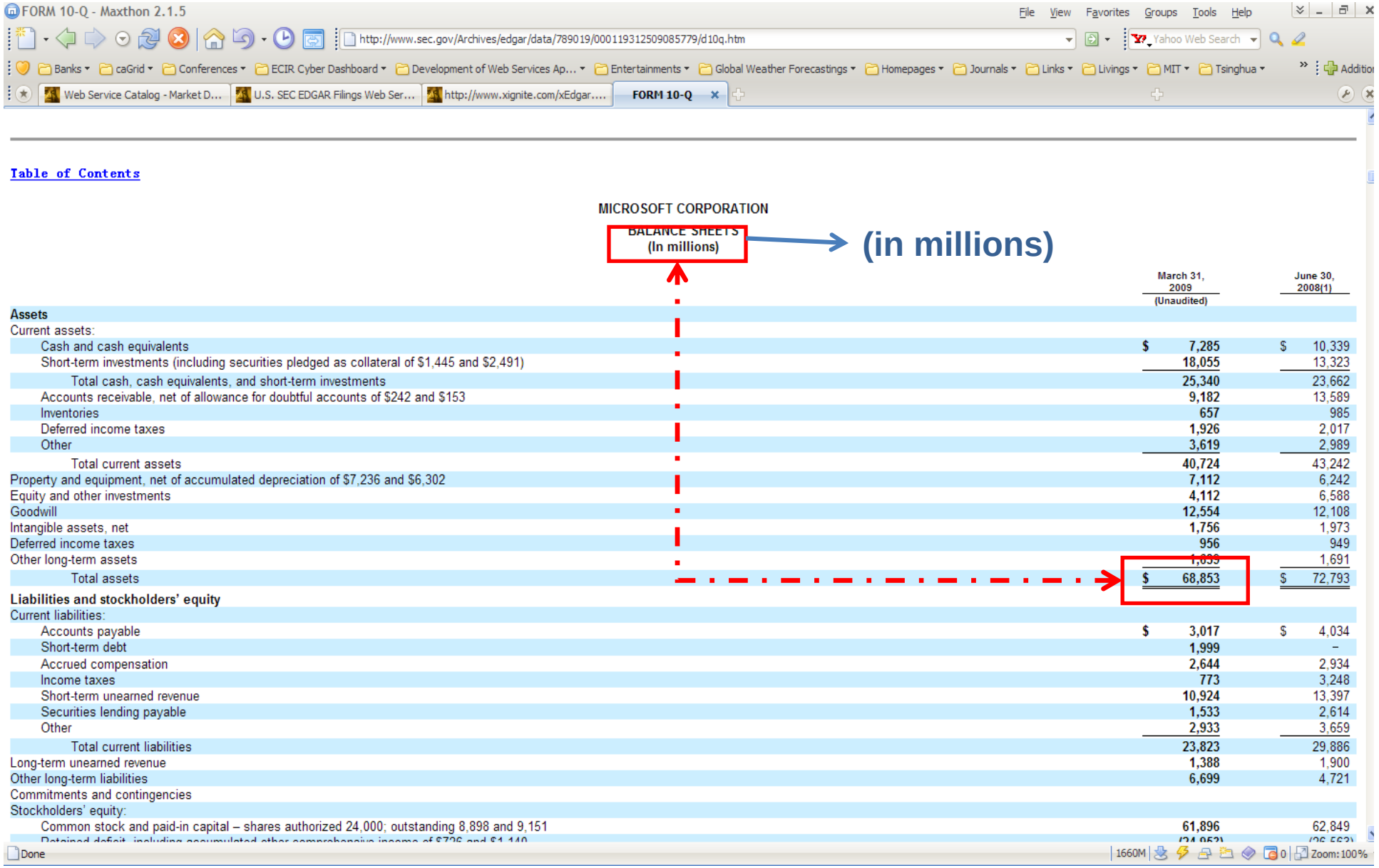
MSFT Total Assets: 68,853

GetTotalAssets for ITWO



MSFT Total Assets: 68,853 vs. ITWO Total Assets: 313,776 ?

(ITWO has Five Times more assets than MSFT?)



	March 31, 2009 (Unaudited)	June 30, 2008(1)
Assets		
Current assets:		
Cash and cash equivalents	$ 7,285	$ 10,339
Short-term investments (including securities pledged as collateral of $1,445 and $2,491)	18,055	13,323
Total cash, cash equivalents, and short-term investments	25,340	23,662
Accounts receivable, net of allowance for doubtful accounts of $242 and $153	9,182	13,589
Inventories	657	985
Deferred income taxes	1,926	2,017
Other	3,619	2,989
Total current assets	40,724	43,242
Property and equipment, net of accumulated depreciation of $7,236 and $6,302	7,112	6,242
Equity and other investments	4,112	6,588
Goodwill	12,554	12,108
Intangible assets, net	1,756	1,973
Deferred income taxes	956	949
Other long-term assets	1,035	1,691
Total assets	$ 68,853	$ 72,793
Liabilities and stockholders' equity		
Current liabilities:		
Accounts payable	$ 3,017	$ 4,034
Short-term debt	1,999	–
Accrued compensation	2,644	2,934
Income taxes	773	3,248
Short-term unearned revenue	10,924	13,397
Securities lending payable	1,533	2,614
Other	2,933	3,659
Total current liabilities	23,823	29,886
Long-term unearned revenue	1,388	1,900
Other long-term liabilities	6,699	4,721
Commitments and contingencies		
Stockholders' equity:		
Common stock and paid-in capital – shares authorized 24,000; outstanding 8,898 and 9,151	61,896	62,849

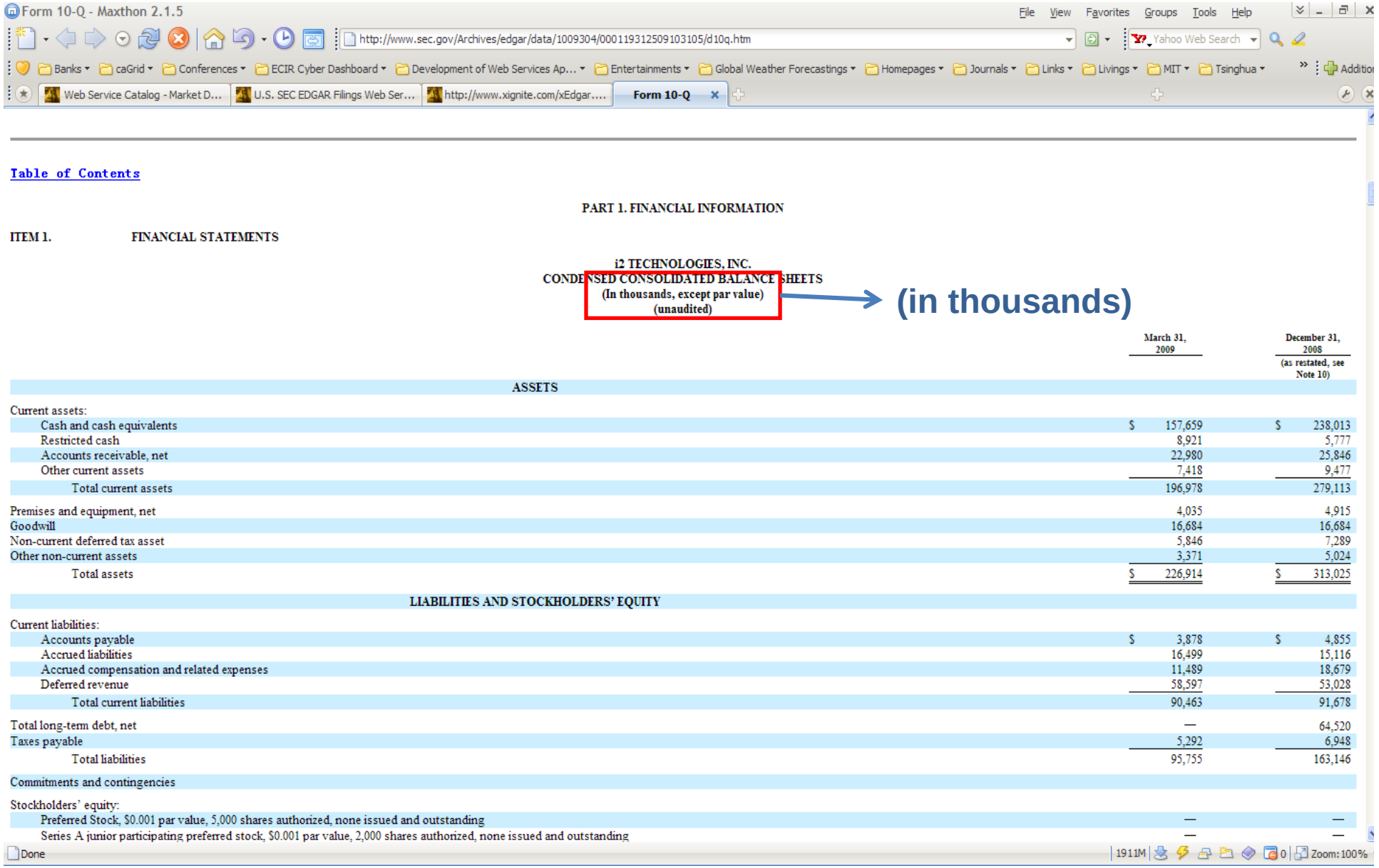


Table of Contents

PART 1. FINANCIAL INFORMATION

ITEM 1. FINANCIAL STATEMENTS

i2 TECHNOLOGIES, INC.
CONDENSED CONSOLIDATED BALANCE SHEETS
(In thousands, except par value)
(unaudited)

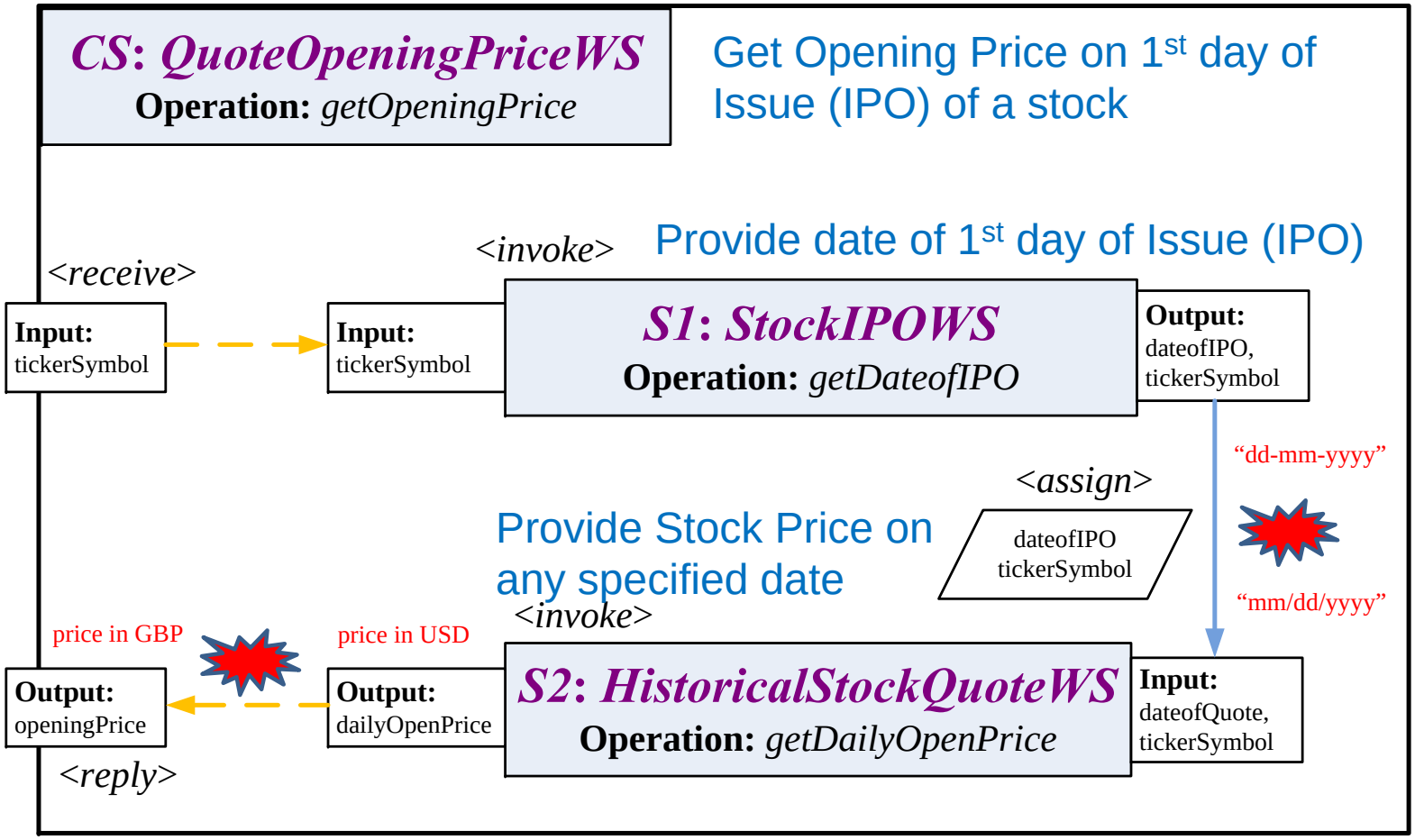
(in thousands)

	March 31, 2009	December 31, 2008 (as restated, see Note 10)
ASSETS		
Current assets:		
Cash and cash equivalents	$ 157,659	$ 238,013
Restricted cash	8,921	5,777
Accounts receivable, net	22,980	25,846
Other current assets	7,418	9,477
Total current assets	196,978	279,113
Premises and equipment, net	4,035	4,915
Goodwill	16,684	16,684
Non-current deferred tax asset	5,846	7,289
Other non-current assets	3,371	5,024
Total assets	$ 226,914	$ 313,025
LIABILITIES AND STOCKHOLDERS' EQUITY		
Current liabilities:		
Accounts payable	$ 3,878	$ 4,855
Accrued liabilities	16,499	15,116
Accrued compensation and related expenses	11,489	18,679
Deferred revenue	58,597	53,028
Total current liabilities	90,463	91,678
Total long-term debt, net	—	64,520
Taxes payable	5,292	6,948
Total liabilities	95,755	163,146
Commitments and contingencies		
Stockholders' equity:		
Preferred Stock, $0.001 par value, 5,000 shares authorized, none issued and outstanding	—	—
Series A junior participating preferred stock, $0.001 par value, 2,000 shares authorized, none issued and outstanding	—	—

Research Motivation

- **Goal of Using Web Services**
  - Web services composition/integration/mashup
    - **Combine multiple web services to create a composite web service**
  - Standards: WSDL (single WS), BPEL (composition process)
- **Long-Standing Challenge for Data Quality**
  - Problems of **data misinterpretation**
    - Different Web services have different assumptions about the interpretation of the exchanged data among them in the composition
    - Root cause: inconsistent data representation, unit, precision, scaling, and meaning, etc.

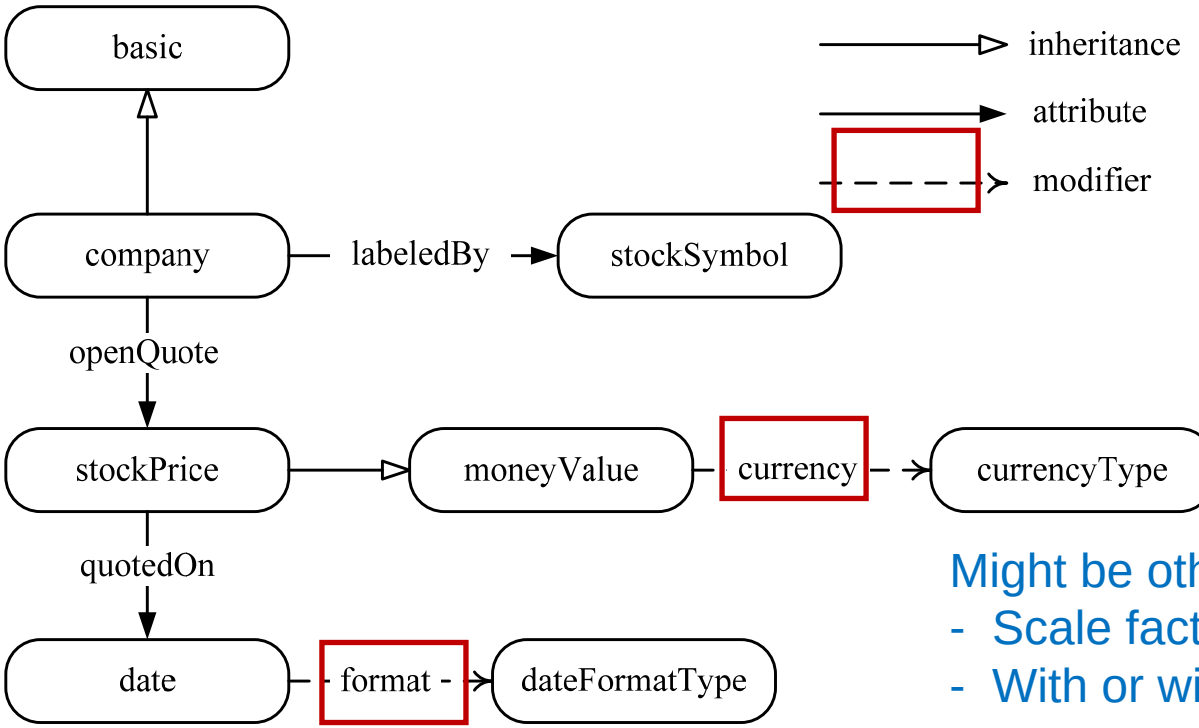
Example Used in our Demonstration: Illustrating Creation of Composite Web Service



Notes: 1. Assumptions of data interpretation are not explicitly represented in WSDLs.
2. Often parallel asynchronous processing (omitted from example)

Context-enriched Ontology Model

- **Ontology with Context Modifiers**



Might be other modifiers, e.g.:
- Scale factor
- With or without fees

Service	Context Definition
CS, S1	*ctxtUK* = {<*format*, dd-mm-yyyy>, <*currency*, GBP>}
S2	*ctxtUS* = {<*format*, mm/dd/yyyy>, <*currency*, USD>}

Specification of Conversion Among Modifier Types (generic conversions – not specific to example)

- **Conversion Representation**

  cvt(*C, m, ctxt_s, ctxt_t, mvs, mvt, vs, vt*)

*cvt*format(date, format, ctxtUK, ctxtUS, "**dd-mm-yyyy**", "**mm/dd/yyyy**", "09-08-2009", *vt*)

vt = "08/09/2009"

- **Conversion Implementation**
  - Xpath functions: *cvt*format

Vt = Concat(substring-before(substring-after(*Vs*,"-") ,"-") ,"/",
substring-before(*Vs*,"-"),"/", substring-after(substring-after(*Vs*,"-") ,"-"))

  - External services: *cvt*currency
    - CurrencyExchangeService

Semantic Annotation of WSDL (using SAWSDL)

```
<wsdl:definitions targetNamespace="http://stockQuote.coin.mit"
  xmlns:stkOntology="http://stockQuote.coin.mit/ontologies/stockOntology#"  xmlns:sawsdl="http://www.w3.org/ns/sawsdl" …
  sawsdl:modelReference="stkOntology#ctxtUK">
<wsdl:types>
<schema elementFormDefault="qualified"  targetNamespace="http://stockQuote.coin.mit"  xmlns="http://www.w3.org/2001/XMLSchema">
  <element name="getDateofIPO">
    <complexType>
      <sequence>
        <element name="tickerSymbol" type="xsd:string“ sawsdl:modelReference="stkOntology#stockSymbol"/>
      </sequence>
    </complexType>
  </element>
  <element name="getDateofIPOResponse">
    <complexType>
      <sequence>
        <element name="getDateofIPOReturn" type="impl:IPOBean"/>
      </sequence>
    </complexType>
  </element>
  <complexType name="IPOBean">
    <sequence>
      <element name="dateofIPO" nillable="true" type="xsd:string“ sawsdl:modelReference="stkOntology#date stkOntology#ctxtUK"/>
      <element name="tickerSymbol" nillable="true" type="xsd:string“  sawsdl:modelReference="stkOntology#stockSymbol"/>
    </sequence>
  </complexType>
</schema>
</wsdl:type>
```

Both **global** and **local** context specification possible.

Reconciliation Approach for Any Web Service Composition

Generic Ontology/SAWSDL context definitions of web services and Conversion Specifications set up

Our Reconciliation Approach

1. **Translating WSDL/BPEL to LOTOS NT**
2. **Detecting Context Conflicts**
3. **Incorporating Conversions into LOTOS NT**
4. **Generating Mediated BPEL**

1. Translating WSDL/BPEL to LOTOS NT

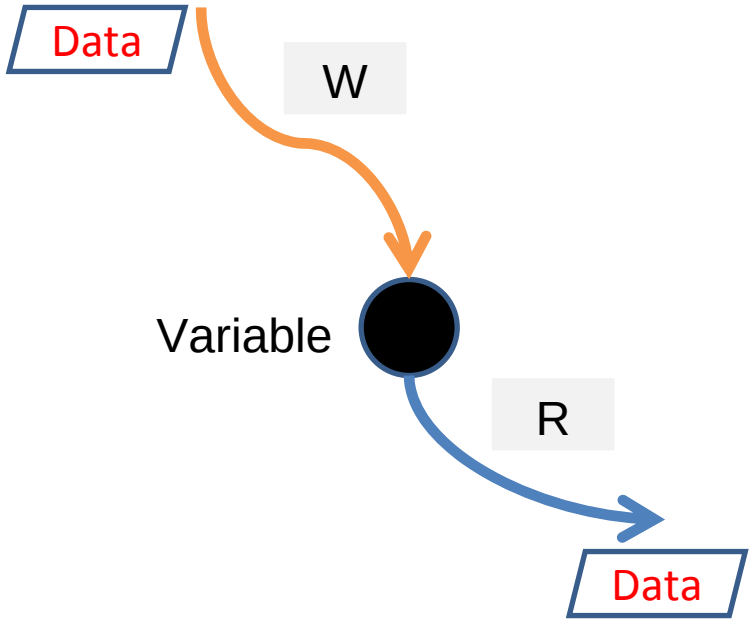
- **Benefits of LOTOS NT**
  - Formalism for verifying composition processes, e.g., deadlock-freeness
  - Independence of any composition languages, e.g., BPEL, OWL-S process
- **Translating Static Aspect to Types**
  - WSDL: data types, messages, operations
  - BPEL: variables, partner links

```
(* @ tickerSymbol:mdlRef="stkOntology#stockSymbol"
 * @ dateofIPO:mdlRef="stkOntology#date stkOntology#ctxtUK"
 *)
type IPOBeanComplexType is
    IPOBeanComplexType (tickerSymbol:string, dateofIPO:string)
end type
```

2. Detecting Context Conflicts

- **2.1 Identify Data Transfers**
  - Explicit data transfers: <*assign*> (BPEL), ":=" (LOTOS NT)
  - Implicit data transfers through variables in the process

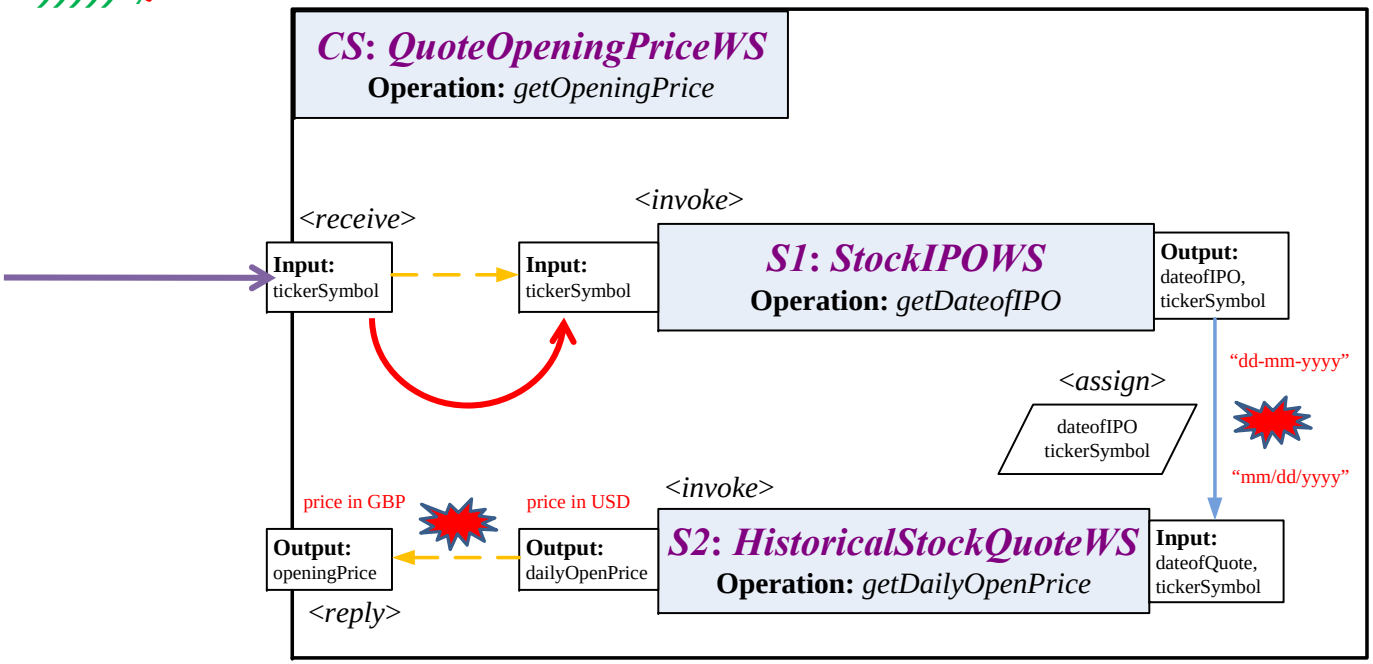
Implicit data transfers refer to the situations that *a piece of data is written to a variable and then directly read from the variable, without assignments between the write and read operations*.



2. Detecting Context Conflicts

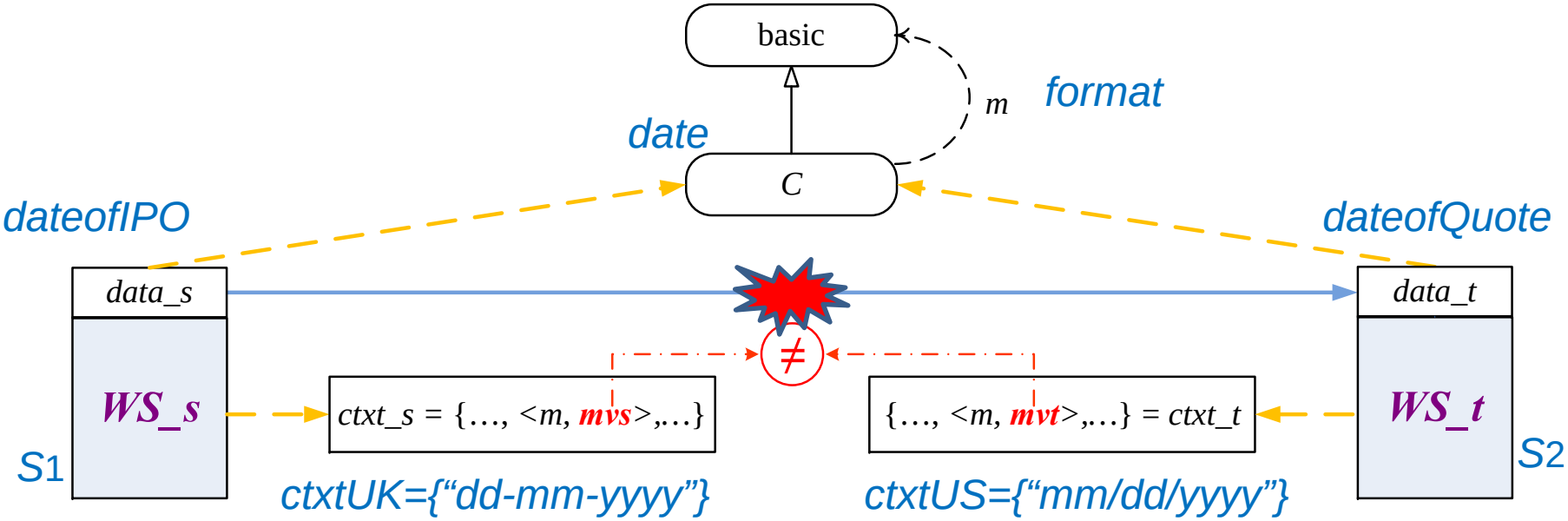
- **Implicit Data Transfers**

QuoteOpeningPricePL(?quoteOpeningPriceRole
(getOpeningPriceInput(getOpeningPriceRequestMessage
(getDateofIPO)))) ;
StockIPOPL(!stockIPOServiceRole(getDateofIPOInput
(getDateofIPORequestMessage(getDateofIPOType
(getDateofIPO))))) ;



2. Detecting Context Conflicts

- **2.2 Examine Each Data Transfer for Conflicts**



Typical Scenario of Context Conflicts within a Data Transfer

2. Detecting Context Conflicts

- **Detected Context Conflicts**

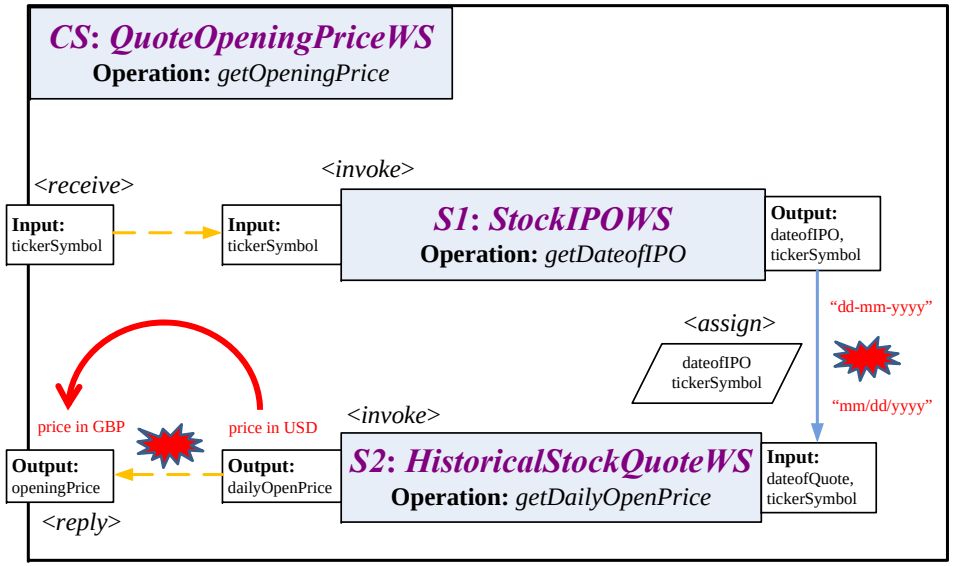
	Source	Target
LOTOS NT	dateofIPO	dateofQuote
Web service	S1	S2
Context	ctxtUK	ctxtUS
date.format	“dd-mm-yyyy”	“mm/dd/yyyy”

	Source	Target
LOTOS NT	getDailyOpenPriceResponse	getDailyOpenPriceResponse
Web service	S2	CS
Context	ctxtUS	ctxtUK
moneyValue.currency	“USD”	“GBP”

3. Incorporating Conversions into LOTOS NT

- **3.1 Make implicit data transfers with conflicts explicit**

HistoricalStockQuotePL (?historicalStockQuoteServiceRole
(getDailyOpenPriceOutput(getDailyOpenPriceResponseMessage
(getDailyOpenPriceResponse)))) ;
QuoteOpeningPricePL(!quoteOpeningPriceRole
(getOpeningPriceOutput(getOpeningPriceResponseMessage
(getDailyOpenPriceResponse))))



3. Incorporating Conversions into LOTOS NT

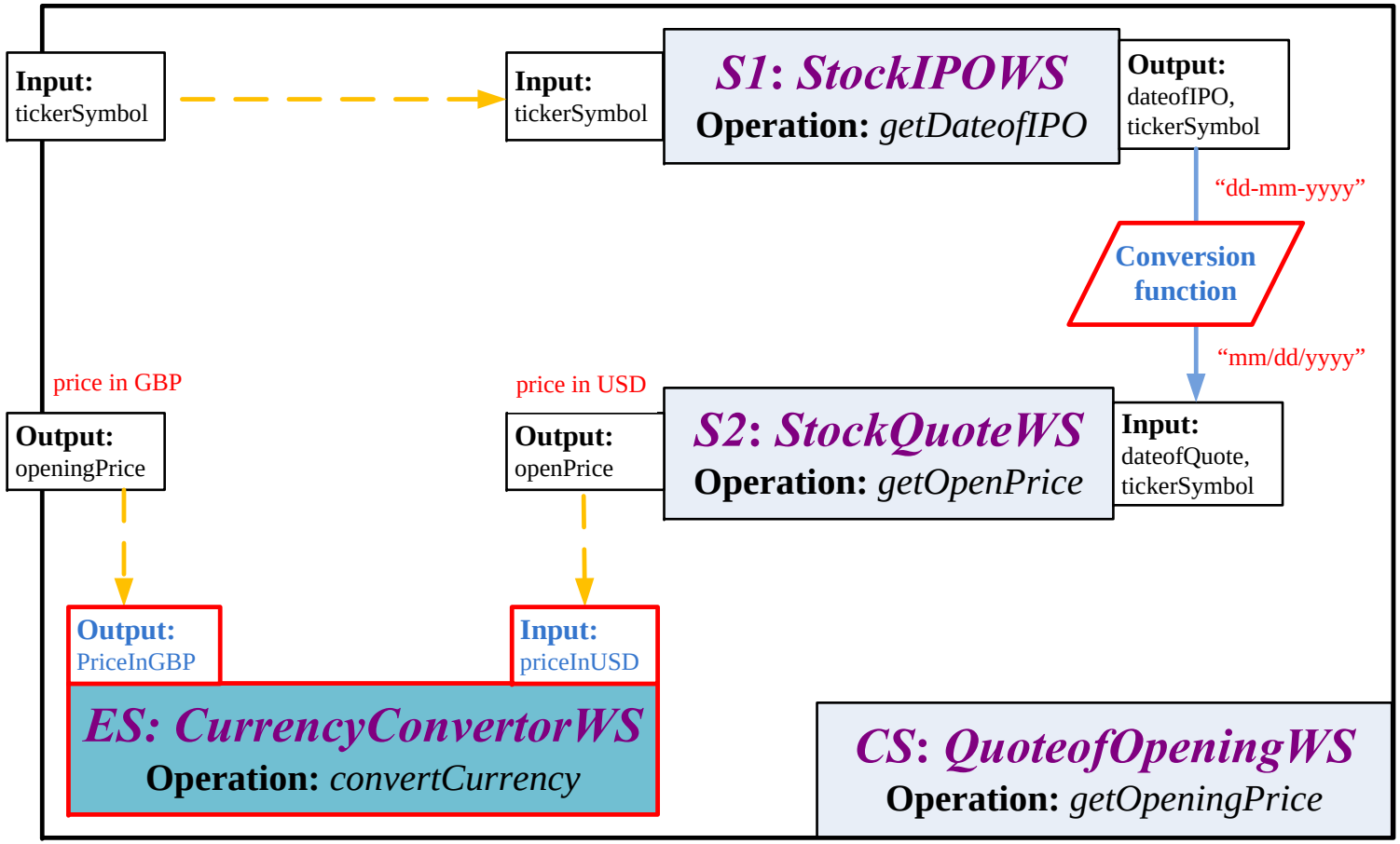
HistoricalStockQuotePL (?historicalStockQuoteServiceRole
(getDailyOpenPriceOutput(getDailyOpenPriceResponseMessage
(getDailyOpenPriceResponse)))) ;
QuoteOpeningPricePL(!quoteOpeningPriceRole
(getOpeningPriceOutput(getOpeningPriceResponseMessage
(getDailyOpenPriceResponse))))

var getOpeningPriceResponse

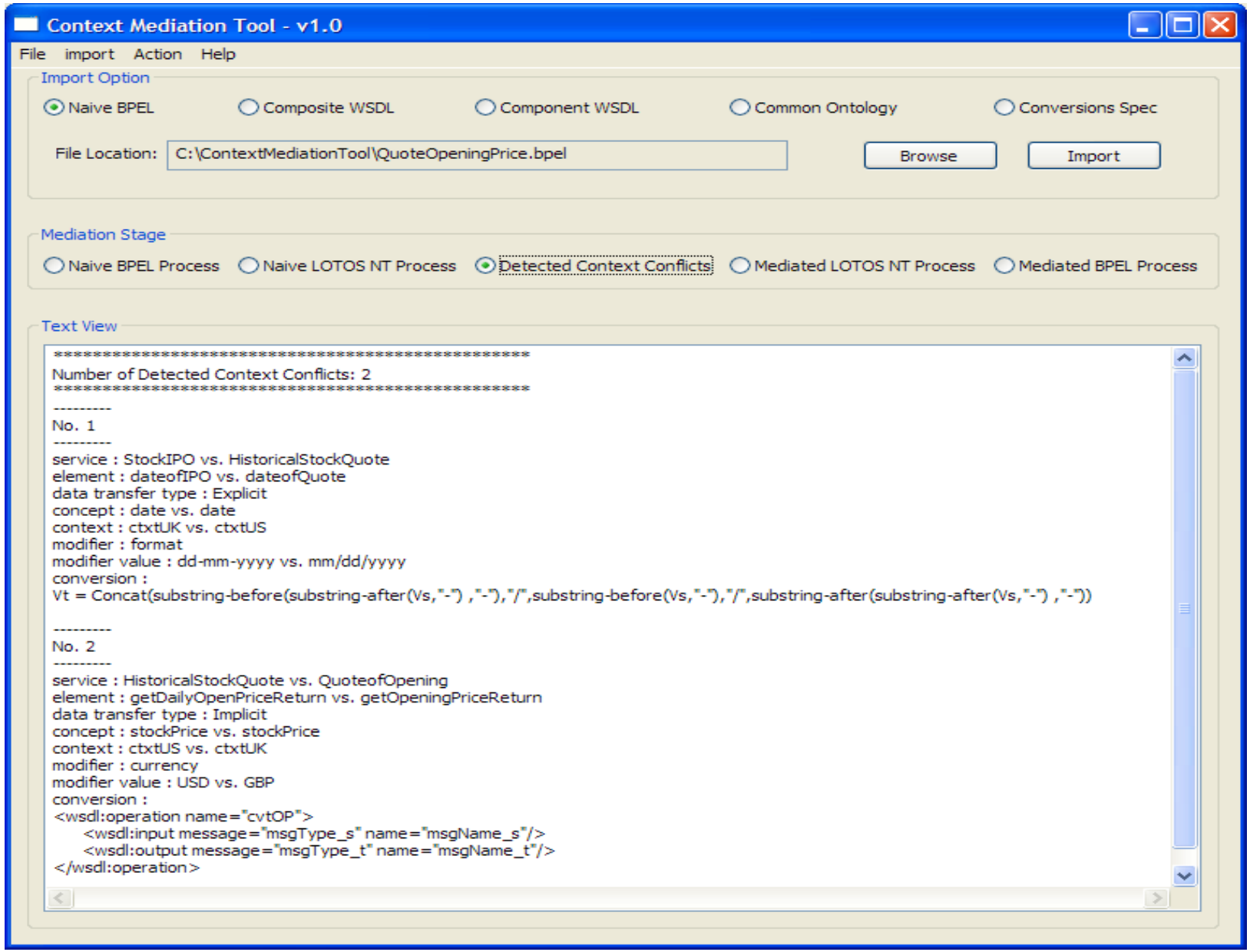
HistoricalStockQuotePL (?historicalStockQuoteServiceRole
(getDailyOpenPriceOutput(getDailyOpenPriceResponseMessage
(getDailyOpenPriceResponse)))) ;
getOpeningPriceResponse := getDailyOpenPriceResponse;
QuoteOpeningPricePL(!quoteOpeningPriceRole
(getOpeningPriceOutput(getOpeningPriceResponseMessage
(getOpeningPriceResponse))))

3. Incorporating Conversions into LOTOS NT

- **3.2 Incorporate conversions and generate mediated process**



Prototype: *CMT* – Illustrating Conflict Detection



Summary and Future Work

- **Strengths**
  - Address data misinterpretation problems to improve quality
  - Separation of concern between service descriptions (e.g., BPEL) and ontology/context descriptions
  - Use the standard-compliant method (i.e., SAWSDL) for semantic annotation
  - Use LOTOS, a kind of process algebra, which can be integrated with behavioral compatibility verification
  - Automates the incorporation of conversions into composition (can assemble complex conversions)

- **Ongoing Work**
  - Extend to deal with semantic heterogeneity at both schematic and data instance level
  - (Semi-)automatic construction of ontology, contexts and annotations

Maybe Web Services were in use during the Napoleonic Wars? *(Data Quality Consequences)*

In 1805, the Austrian and Russian Emperors agreed to join forces against Napoleon. The Russians promised that their forces would be in the field in Bavaria by **Oct. 20**.

The Austrian staff planned its campaign based on that date in the **Gregorian calendar**. Russia, however, still used the ancient **Julian calendar**, which lagged 10 days behind.

The calendar difference allowed Napoleon to surround Austrian General Mack's army at Ulm and force its surrender on Oct. 21, well before the Russian forces could reach him, ultimately setting the stage for Austerlitz.

Source: David Chandler, *The Campaigns of Napoleon*, New York: MacMillan 1966, pg. 390.

Thanks again for your interest in this work.
Please refer to the paper for more details.

Any questions?

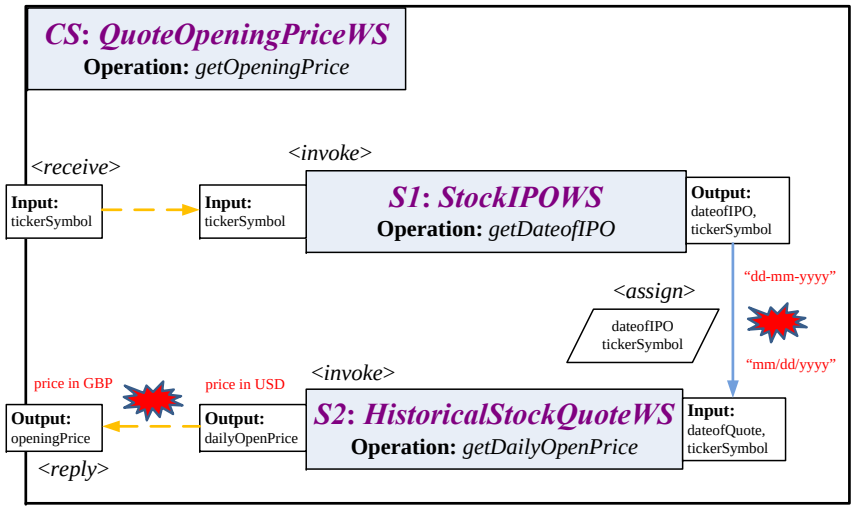
Backup Slides

1. Translating WSDL/BPEL to LOTOS NT

- **Translating Dynamic Aspect to Process**

BPEL	LOTOS NT
<receive variable=”v” …/>	(*c*?*v*)
<reply variable=”v” …/>	(*c*!*v*)
<invoke inputVariable=”v1” *outputVariable=”v2”…/>*	(*c*!*v1* ; *c*?*v2*)
<assign …> *<copy>* *<from variable=”v1”>* *<to variable=”v2”/>* *</copy>* *<copy>* *<from variable=”v3”>* *<to variable=”v4”/>* *</copy>* *</assign>*	(*v2* := *v1*; *v4* := *v3*)
<sequence …> *<…activity1…/>* *<…activity2…/>* *</sequence>*	(*action1* ; *action2*)

1. Translating WSDL/BPEL to LOTOS NT



QuoteOpeningPricePL(?quoteOpeningPriceRole (getOpeningPriceInput(getOpeningPriceRequestMessage (getDateofIPO)))) ;
StockIPOPL(!stockIPOServiceRole(getDateofIPOInput (getDateofIPORequestMessage(getDateofIPOType (getDateofIPO))))) ;
StockIPOPL(?stockIPOServiceRole(getDateofIPOOutput (getDateofIPOResponseMessage(getDateofIPOResponse)))) ;
getDailyOpenPrice := getDailyOpenPriceType (getDateofIPOResponse.getDateofIPOReturn.tickerSymbol, getDateofIPOResponse.getDateofIPOReturn.dateofIPO) ;
HistoricalStockQuotePL(!historicalStockQuoteServiceRole (getDailyOpenPriceInput(getDailyOpenPriceRequestMessage (getDailyOpenPrice)))) ;
HistoricalStockQuotePL (?historicalStockQuoteServiceRole (getDailyOpenPriceOutput(getDailyOpenPriceResponseMessage (getDailyOpenPriceResponse)))) ;
QuoteOpeningPricePL(!quoteOpeningPriceRole (getOpeningPriceOutput(getOpeningPriceResponseMessage (getDailyOpenPriceResponse))))