

On Analyzing and Specifying Concerns for Data as a Service

Hong-Linh Truong and Schahram Dustdar

Distributed Systems Group
Vienna University of Technology

truong@infosys.tuwien.ac.at
<http://www.infosys.tuwien.ac.at/Staff/truong/>

Acknowledgment: Marco Comerio, G.R. Gangadharan, Roman Khazankin, Reinhard Pichler, Andrea Maurino, Vadim Savenkov,



2009 IEEE Asia-Pacific Services Computing Conference
(IEEE APSCC 2009)

December 7-11, 2009

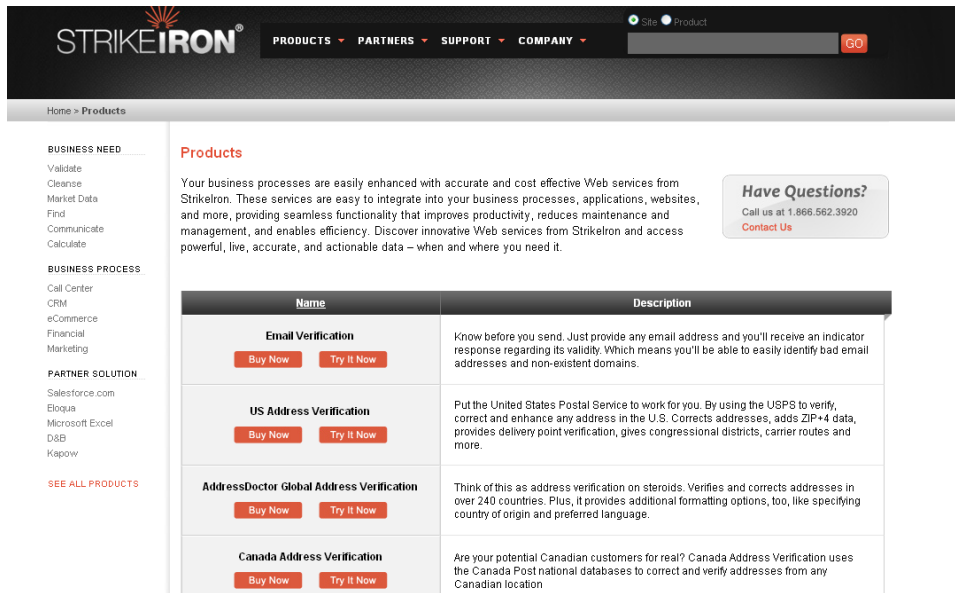
Biopolis, Singapore

- Background and motivation
- DaaS concerns
- Specifying DaaS concerns
- Linking DaaS concerns to services
- Current prototype
- Some studies of DaaS concerns in current service descriptions
- Conclusion and future work

Background

- Web services technologies, the SaaS model and the cloud computing model foster the concept of data/information as a service (DaaS)
- No precise definition but DaaS
 - *Provide data capabilities* rather than provide computation on data or data based on computation
- Providing DaaS is an increasing trend
 - In both business and e-science environments
 - Bio data, weather data, company balance sheets, etc., via Web services
 - Academic research and industrial relevant research topics

- Read-only DaaS versus CRUD DaaS
- Service APIs versus Data
 - *Service APIs* are used to *CRUD data*
 - They are not the same wrt concerns



STRIKEIRON PRODUCTS PARTNERS SUPPORT COMPANY

Home > Products

BUSINESS NEED
 Validate
 Cleanse
 Market Data
 Find
 Communicate
 Calculate

BUSINESS PROCESS
 Call Center
 CRM
 eCommerce
 Financial
 Marketing

PARTNER SOLUTION
 Salesforce.com
 Eloqua
 Microsoft Excel
 D&B
 Kapow

SEE ALL PRODUCTS

Products

Your business processes are easily enhanced with accurate and cost effective Web services from StrikeIron. These services are easy to integrate into your business processes, applications, websites, and more, providing seamless functionality that improves productivity, reduces maintenance and management, and enables efficiency. Discover innovative Web services from StrikeIron and access powerful, live, accurate, and actionable data – when and where you need it.

Have Questions?
 Call us at 1.866.562.3920
[Contact Us](#)

Name	Description
Email Verification Buy Now Try It Now	Know before you send. Just provide any email address and you'll receive an indicator response regarding its validity. Which means you'll be able to easily identify bad email addresses and non-existent domains.
US Address Verification Buy Now Try It Now	Put the United States Postal Service to work for you. By using the USPS to verify, correct and enhance any address in the U.S. Corrects addresses, adds ZIP+4 data, provides delivery point verification, gives congressional districts, carrier routes and more.
AddressDoctor Global Address Verification Buy Now Try It Now	Think of this as address verification on steroids. Verifies and corrects addresses in over 240 countries. Plus, it provides additional formatting options, too, like specifying country of origin and preferred language.
Canada Address Verification Buy Now Try It Now	Are your potential Canadian customers for real? Canada Address Verification uses the Canada Post national databases to correct and verify addresses from any Canadian location.



Infochimps
 Find the world's data

Explore > Datasets Collections Tags sell share search

Discover, share and sell data of any size, topic, or format.

Open source knowledge. Anyone can post data under an open license for the world to share and edit, forever, for free. [Signup](#) to share data! [Sign up](#) as a user to edit or share data. [Learn more...](#)

The first open marketplace for data. For anything from polling surveys to market research to fantasy sports statistics, we can connect your data to a massive audience of customers. You control the terms, you set the price, we handle storage, distribution and billing. Sell data now! [Vendors](#) Register as a vendor to sell data. [Learn more...](#)

Some Interesting Datasets

- Area Code and Exchange to Location, North America (NPA/NOX)
- Number of Governors, by Political Party Affiliation: 1970 to 2007
- Article Search API - NYTimes.com
- Current Cigarette Smoking by Sex and State: 2005
- Measuring Worth: Dollar-Pound Exchange Rate From 1791
- Population and Area: 1790 to 2000
- Expectation of Life and Expected Deaths, by Race, Sex, and Age: 2004
- Word List - 10,000+ Common Place Names

Top Tags

government census population america demographics
 state selected olympics type eutransparency team character
 race finance statistics country industry summary age
 corpus income characteristics number science party language
 sales sex rates school expenditures public federal name
 access-www list employment player retail revenue election
 hispanic origin station political foreign health image ... and many more

Motivation

- Data-specific concerns need for
 - Selecting data services based on provided data and service contracts
 - Evaluating the compatibility of service contracts in data composition
 - Supporting quality-aware data composition from multiple data services
- Data-specific concerns combined with service APIs specific concerns
 - Not just QoS based service selection

Motivation (cont.)

- DaaS are currently considered like any other Web services
 - WSDL/WADL description + QoS + pricing information (mostly in HTML form)
- But concerns on data are different from that on service APIs
- Where are the data-relevant concerns in service descriptions?
 - E.g., data quality, usage permission, and data ownerships
- How data-relevant concerns can be combined with service-relevant concerns?

Existing Work

- QoS description and QoS-based Web services selection are well researched
 - Googling "QoS-based Web services selection,, ~ 20.000
- Data Quality is well-known in database community
 - E.g., see ACM Computing Survey 41(3):2009 on data qualities done by Batini et al.
- (Service) Licensing is currently being studied for SaaS
- Several licenses for data are introduced but in human-readable form only
 - E.g., Talis community license, the Open Knowledge Foundation Wiki, the Open Database License
- Intensive discussion on laws and regulations on cloud computing
 - E.g., see Davide Maria Parrilli's work
- Data Governance: e.g., see the IBM data governance maturity model

Issues and Approach

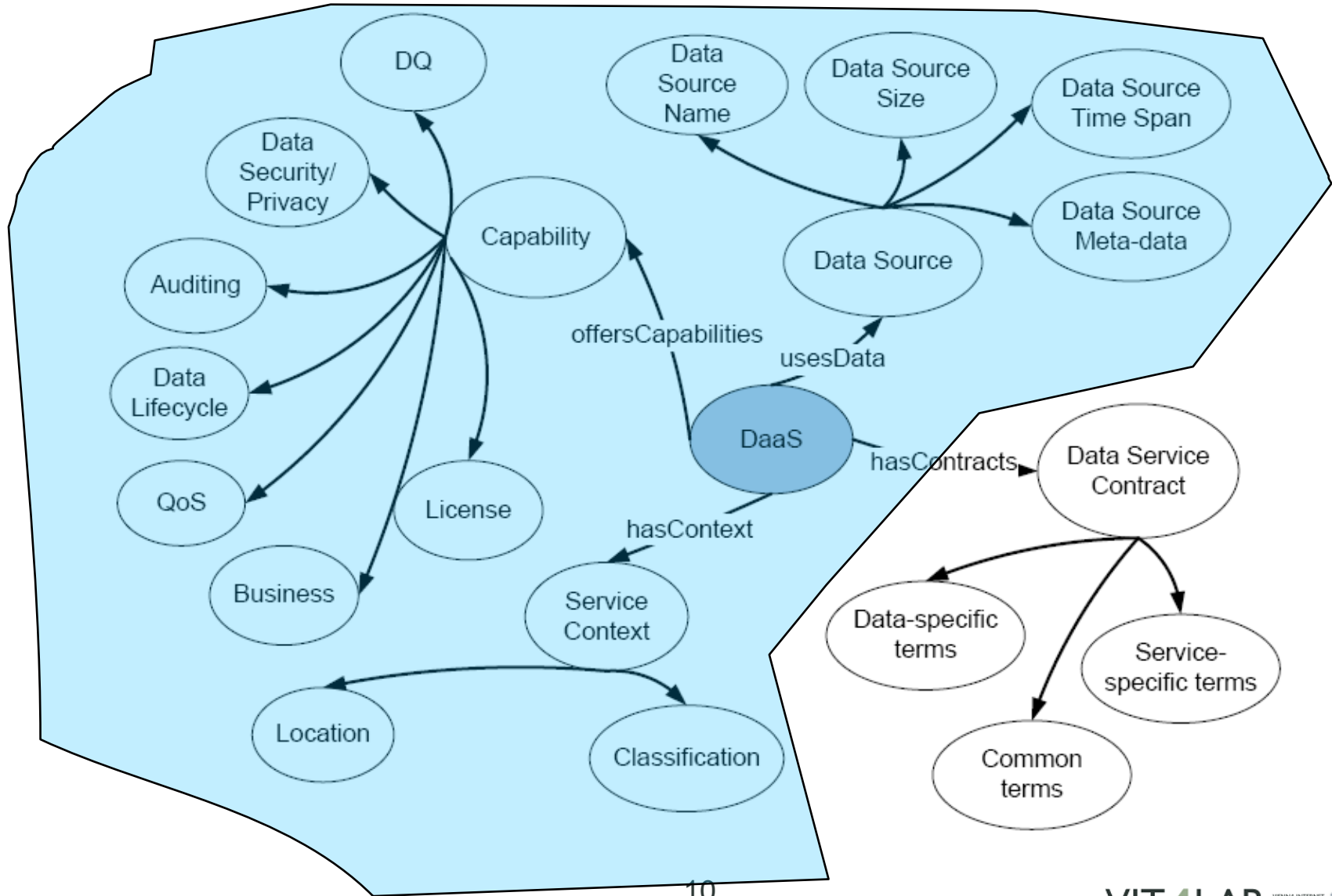
- Issues
 - DaaS concerns include QoS, DQ, service licensing, data licensing, data governance, etc.
 - There is a lack of techniques for the publishing, discovery, selection and evaluation of data concerns
 - There is a lack of techniques for integrating concerns for DaaS
 - Data concerns and Service APIs concerns
- This talk focuses on publishing information that characterizes DaaS
 - What are main DaaS concerns (non-functional parameters) and how to specify them and provide them for the data service selection and contract compatibility?
 - Some empirical studies on existing DaaS descriptions
- We are not talking about how to evaluate concerns and monitor them

The Importance of Concerns in Data Consumer's View

Concerns	Read-only DaaS	CRUD DaaS
Data Quality	Important factor for the selection of DaaS. For example, the accuracy and completeness of the data, whether the data is up-to-date	Expected some support to control the quality of the data in case the data is offered to other consumers
Data source	Important factor for the trustworthiness of the DaaS.	
Data & Service Usage	Important factor, in particular, price, data and service APIs licensing, law enforcement, and IPRs	Important factor, in particular, price, service APIs licensing, and law enforcement
Data Governance		Important factor, for example, the security and privacy compliance, data distribution, and auditing
QoS	Important factor, in particular availability and response time	Important factor, in particular, availability, response time, dependability, and security
Service Context	Useful factor, such as classification and service type (REST, SOAP), location	Important factor, e.g. location (for regulation compliance) and versioning



Conceptual Model for DaaS Concerns and Contracts



Capability Concerns

- Data Quality capabilities
 - Based on well-established research on data quality
 - Timeliness, uptodate, free-of-error, cleaning, consistency, completeness, domain-specific metrics, etc.
 - We mainly support the specification of DQ metrics for the whole DaaS but possible to extend to the service operation level
- Data Security/Privacy capabilities
 - Data protection within DaaS, e.g. encryption, sensitive data filtering, and data privacy
 - Many terms are based on the W3C P3P

Capability Concerns (cont.)

- Auditing capabilities
 - Logging, reporting (e.g., daily, weekly, and monthly), and warning
 - Support system maintenance, SLA monitoring, billing, and taxation
- Data lifecycle
 - Backup/recovery, distribution (e.g., a service is in Europe but data is stored in US), and disposition
 - Support system maintenance but also regulation on data

Capability Concerns (cont.)

- QoS capabilities are applied to service APIs
 - Based on well-researched QoS for Web services
 - Performance capabilities
 - e.g., latency, response time and throughput
 - Dependability capabilities,
 - e.g., availability, reliability, accessibility, security
- Business
 - *Pricing model* (flat rate, pay-per-use, with/without transaction conditions) and *Price*
 - Service credit for reward or compensation
 - e.g. Amazon service credits

Capability Concerns (cont.)

- Data and service license
 - Usage permission: for data (distribution, transfer, personal use, etc.) and for service APIs (adaptation, composition, derivation, etc.)
 - We utilize some terms from ODRL/ODRL-S
 - Copyrights
 - Liability: e.g., who is responsible for the loss due to a network disruption?
 - Law enforcement (e.g., US or European court)
 - Domain specific IRPs

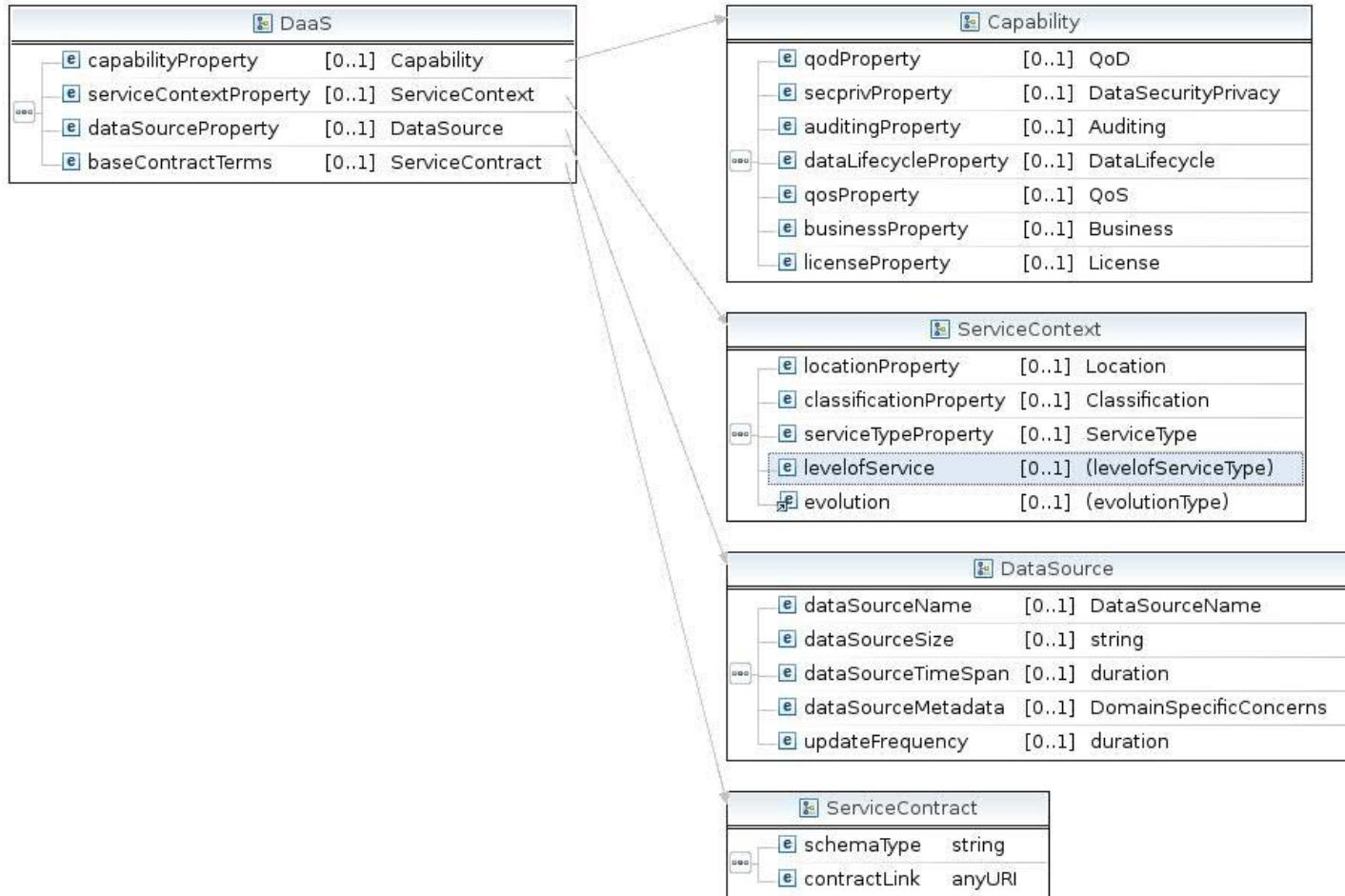
Data Source Concerns

- A DaaS may utilize data from many sources.
- Similar DaaSs may utilize data from the same source
- Data source properties
 - Name: e.g. ddfFlus or DataFlux or Mr A
 - Size
 - Timespan: the duration of collected data, e.g., more than 4 years in the eBay Data License
 - Update Frequency: how often the data is updated
 - Etc.

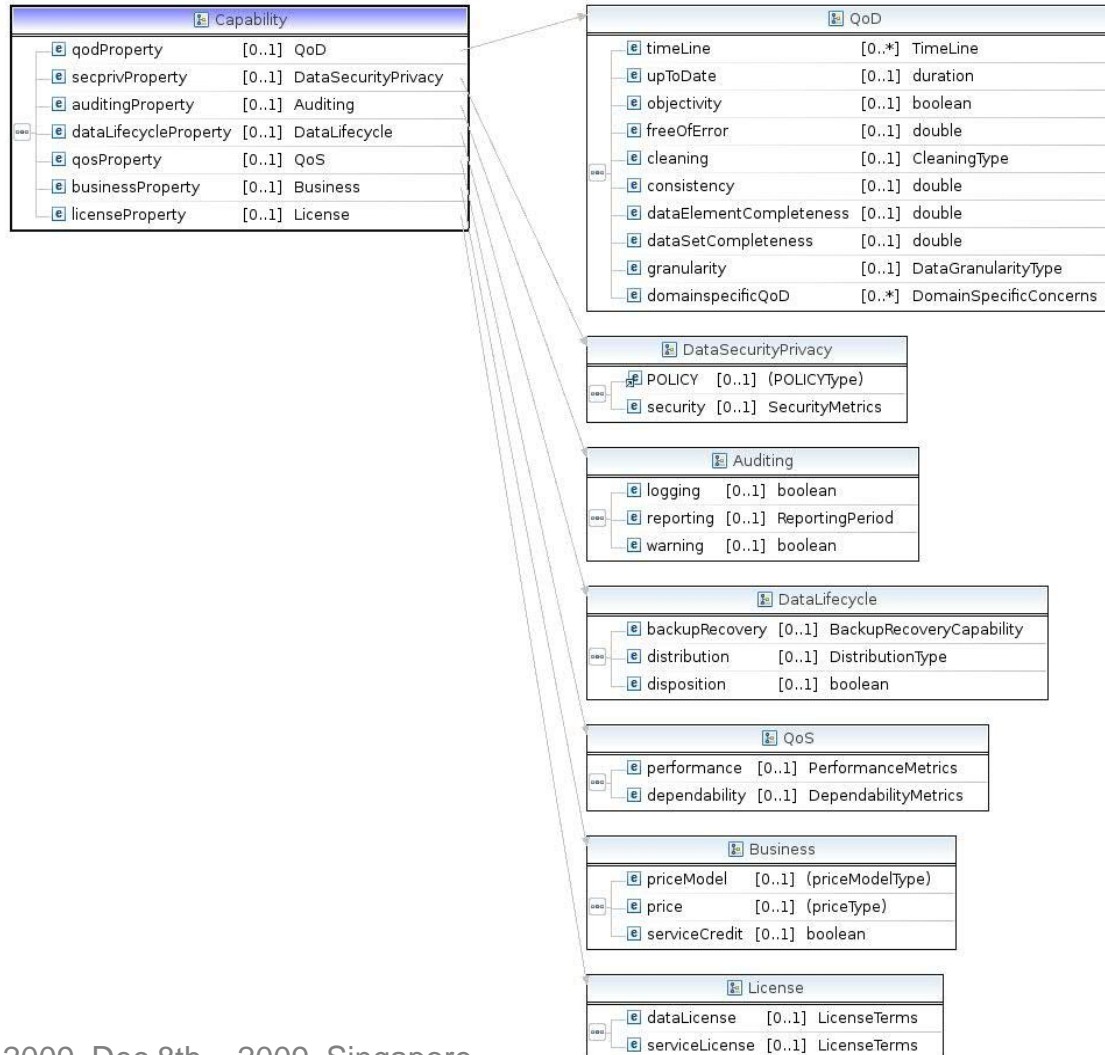
Service Context Concerns

- Location:
 - Selecting a DaaS in Amazon US Zone or European Zone?
- Service Type: REST or SOAP?
 - E.g., mobile client daas
- Level of Service
- Service Classification
 - Based on UNSPSC Code Classification Services
- Data Classification
- Service/data versioning

XML Diagram for DaaS Specification

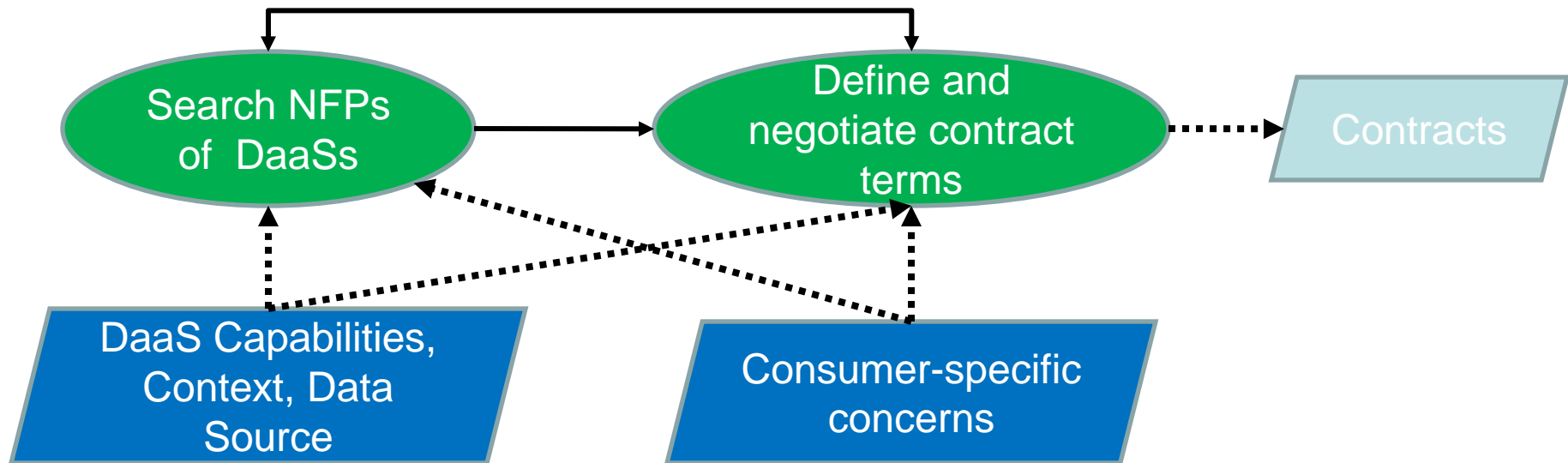


XML Diagram for the DaaS Capability Specification



From Capability/Context to Service Contract

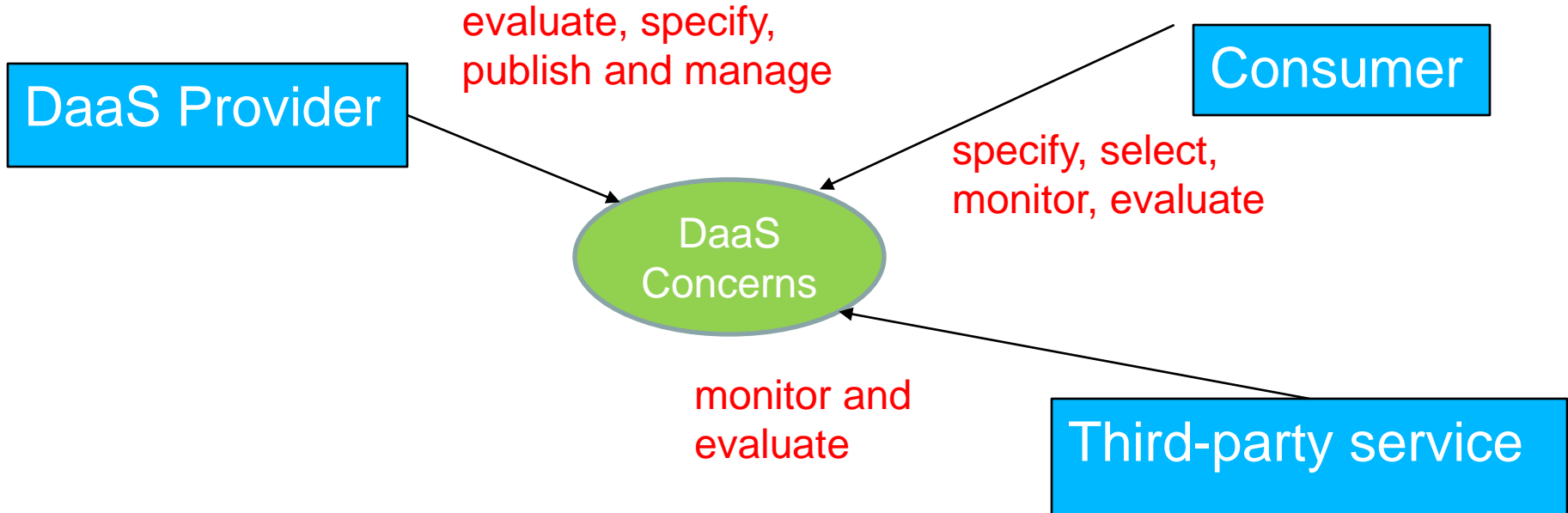
Non-functional parameters (NFPs) to Service Contracts



- A service contract includes a set of generic, data-specific and service-specific conditions established based on concerns

Populating DaaS Concerns

The role of stakeholders in the most trivial view



- We address the specification, publishing and management of DaaS concerns
 - To support the selection of DaaSs
- Monitoring and evaluation are currently open

Implementation

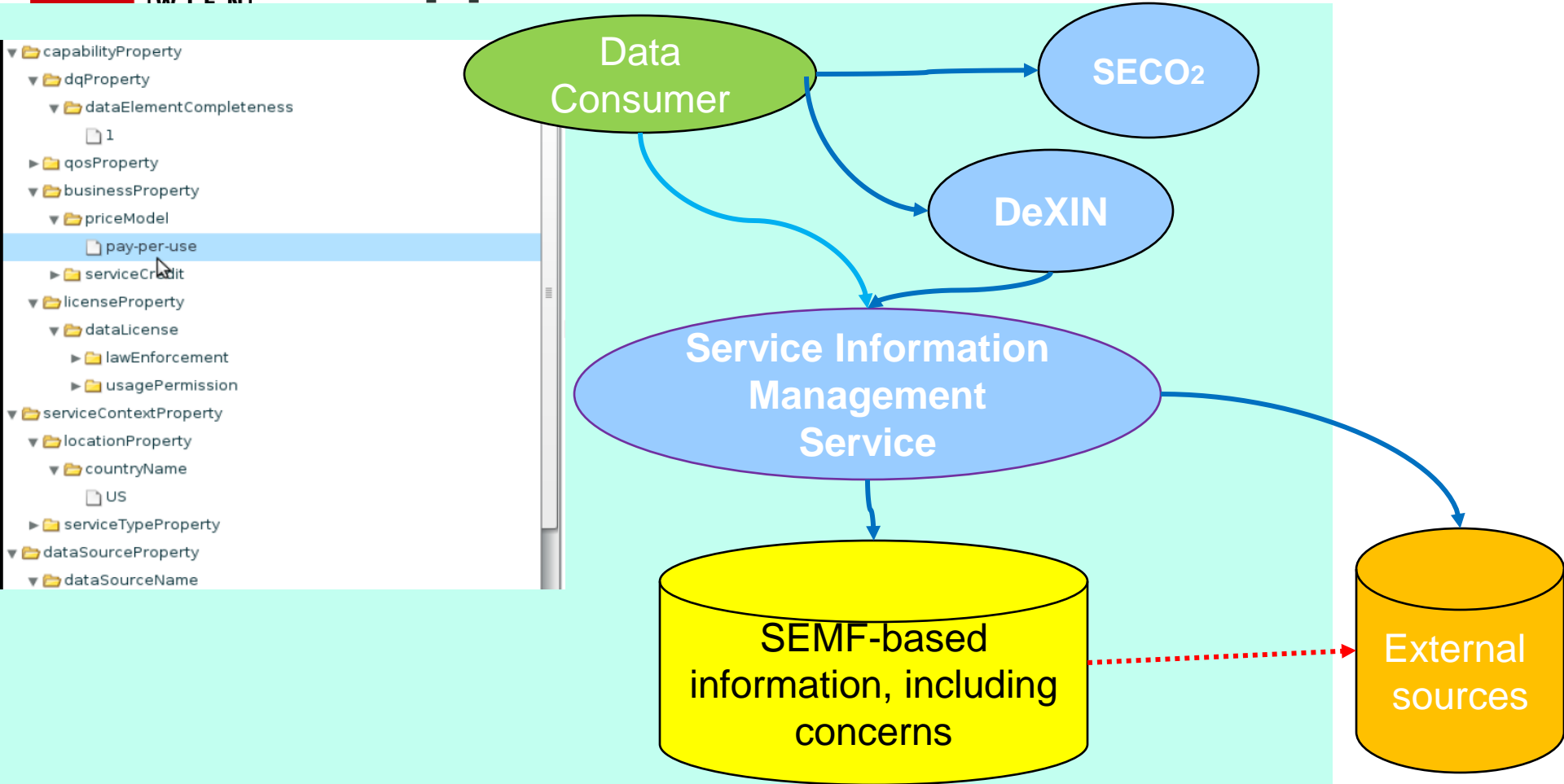
- Concern specifications
 - Possible solutions: XML, RDF, and OWL
 - Our implementation is based on XML/RDF
 - Easy to reuse vocabularies defined in other standards
 - Link to external domain-specific models of concerns using URIs
- Publishing and linking concerns to services
 - Possible solutions: annotating WSDL, SAWSDL, and external management services
 - We use our SEMF model. Concerns are managed via services supporting the evolutionary management

Example of linking concerns with other type of data

Based on SEMF (Service Evolution Management Framework) [SEAA 08]

```
<title>CorteraCreditPulseService</title>
<entry>
  <title>Interface</title>
  <summary>WSDL Interface</summary>
  <category label="Web Service Description" scheme="http://www.dmoz.org/Computers/
    Programming/Internet/Service-Oriented_Architecture/Web_Services/WSDL"
    term="Interface" />
  <content type="application/wsdl+xml" src="http://ws.strikeiron.com/
    CorteraCreditPulse2?WSDL" />
</entry>
<entry>
  <title>DaaS Concerns</title>
  <summary>Data Concerns</summary>
  <category label="Data Concerns" term="DaaSConcern" />
  <content type="application/xml" src="http://www.infosys.tuwien.ac.at/prototyp/SOD1/
    dataconcerns/samples/CorteraCreditPulseConcerns.xml" />
</entry>
```

Support DaaS Concerns Selection



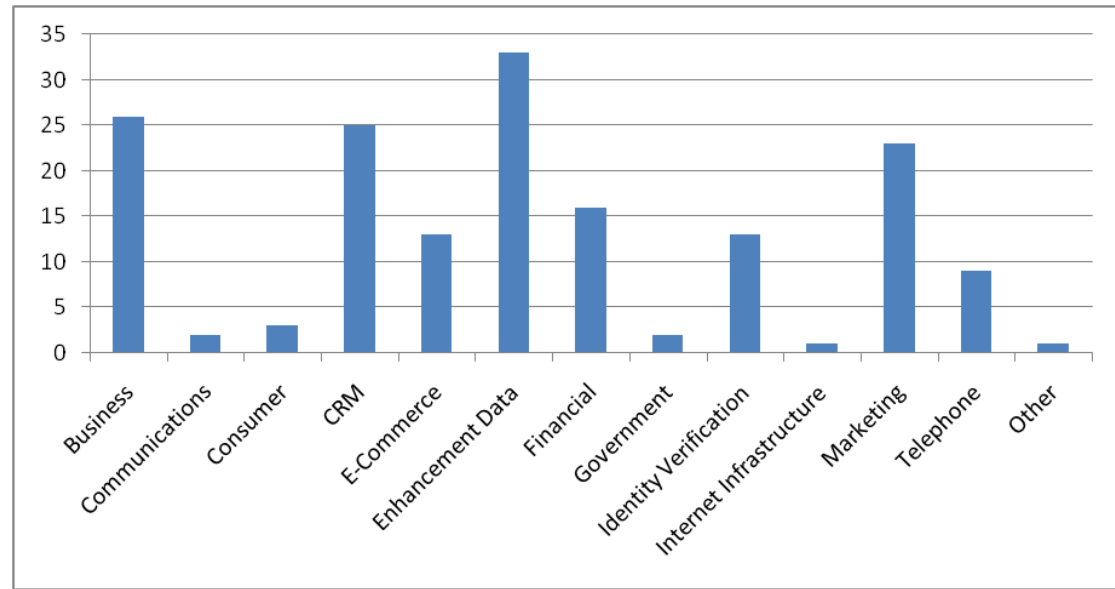
- DeXIN: Distributed XQuery over Heterogeneous Data Sources [ICEIS09, ICWE09]
- SECO2 : Service Contract Compatibility [ICSOC09]

Some Studies

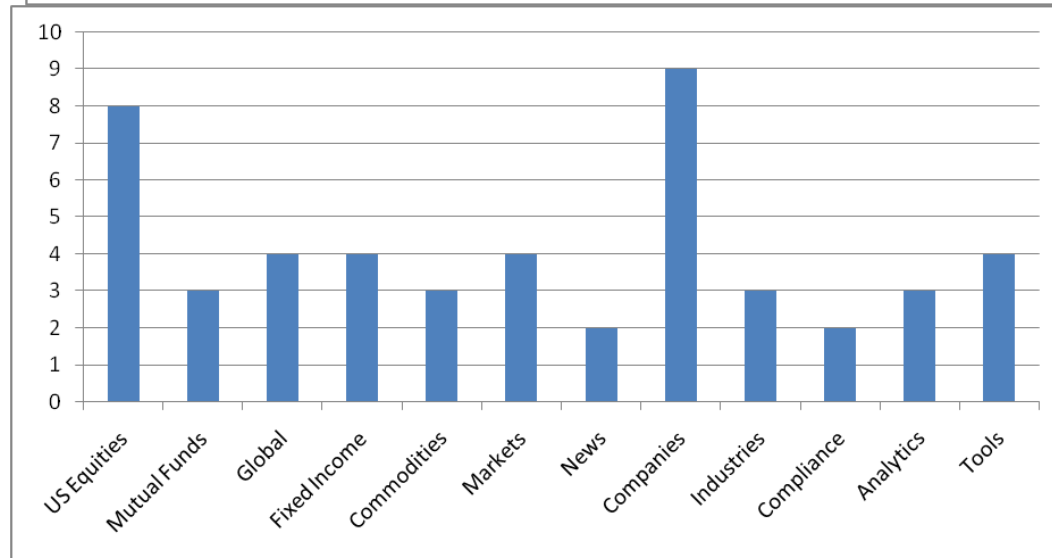
- We are not aware of any provider that publishes DaaS's concerns in a well-defined form
 - Only HTML
- Our studies examines the description of DaaS
 - Enterprising computing
 - Strikelron, Xignite, serviceobjects.NET, WebserviceX, XWebServices, AERS, Amazon
 - E-science
 - GBIF (Global Biodiversity Information Facility), EBI (European Bioinformatics Institute) Web Services, EMBRACE Service Registry, and BioCatalogue

Service Classification

- Strikelron
Web
services

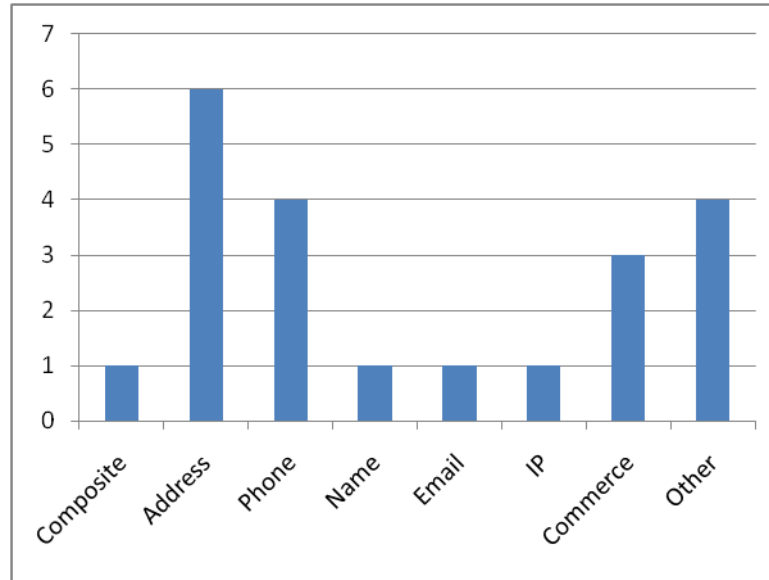


- Xignite
Web
services

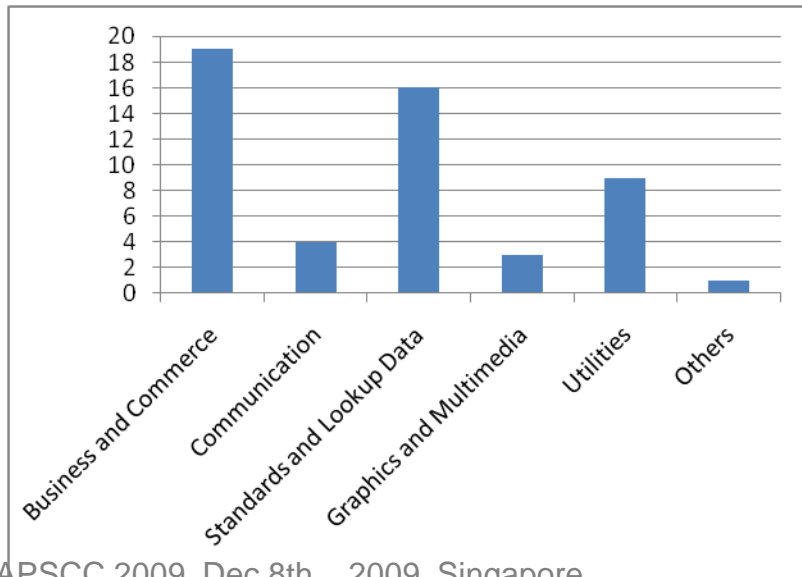


Service Classification

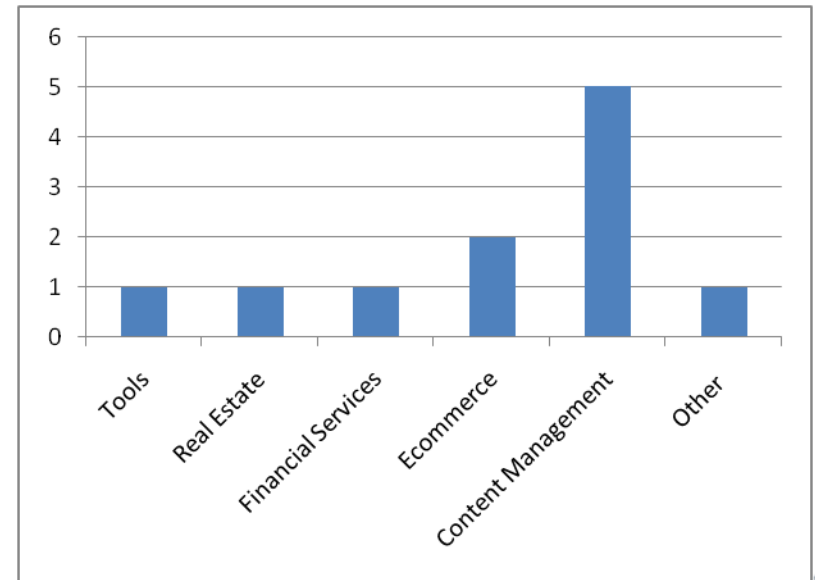
- ServiceObjects
Web Services



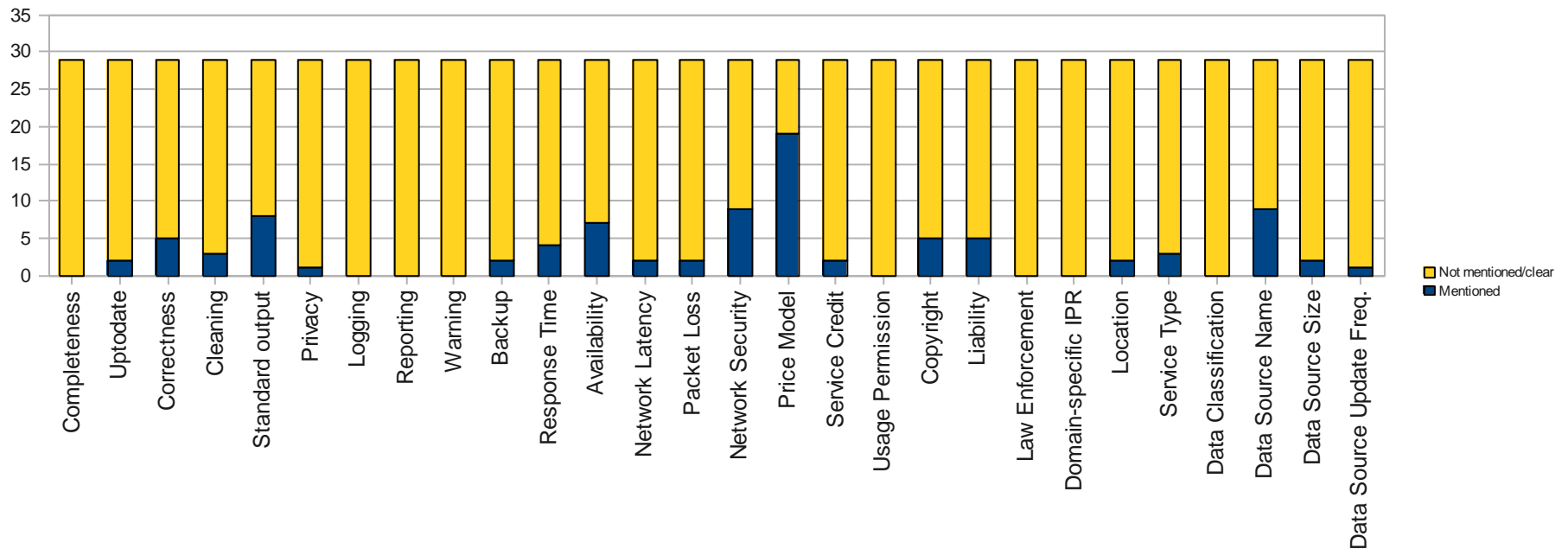
- WebservicesX Web services



- XWebService Web services



- 29 services from 7 providers, most are SOAP-based



From the DaaS description point of view

Service Registries	DQ	QoS	Business	Licensing	
				Ownership	Usage permission
GBIF	No	No	No	unstructured	unstructured
EBI Web Services	No	No	No	No	No
EMBRACE Service Registry	No	No	No	No	No
BioCatalogue	No	No	unstructured	unstructured	unstructured

Conclusion and Future Work

- This paper presents
 - The importance of having DaaS concerns to be explicitly specified in a study of existing concerns
 - A specification and management technique for DaaS concerns
- Future work
 - Enhance empirical studies on current concerns for DaaS
 - Apply DaaS concerns to bioinformatic and biomechanic DaaS
 - Support DaaS concern in data composition/mashup tools and contract compatibility evaluation
 - Develop a service engineering approach for DaaS concerns, and concern monitoring and evaluation
 - Need a joint effort between service engineering and data engineering research

<http://www.infosys.tuwien.ac.at/prototyp/SOD1/>

Thanks for your attention!

Hong-Linh Truong
Distributed Systems Group
Vienna University of Technology

truong@infosys.tuwien.ac.at
<http://www.infosys.tuwien.ac.at/Staff/truong/>

Austria