Privacy-Preserving WebID Analytics on the Decentralized Policy-Aware Social Web

Yuh-Jong Hu
hu@cs.nccu.edu.tw

Emerging Network Technology (ENT) Lab.
Department of Computer Science
National Chengchi University, Taipei, Taiwan

August-11-2014

The 2014 IEEE/WIC/ACM Int. Conf. on Web Intelligence, Warsaw, Poland

Part I

Introduction


Motivations

1. Centralized closed social networking sites are *walled gardens* that limit people activity on a single site.
2. Big data analytics has been proposed for online social networks, but the related privacy protection issue does not arise much attention.
3. Statistical Disclosure Control (SDC) methods have been well-developed for microdata protection, they are also possibly used for data disclosure protection in online social networks.
4. Semantic Web technology has been used for establishing a privacy-aware policy Web architecture to provide flexible and effective privacy-preserving data analytics services.

Research Goals

1. We argue why we choose decentralized but not the centralized online social networking architecture for WebID analytics.
2. How can we provide privacy-preserving *batch* and *interactive* WebID analytics with effective and flexible services for data analysts?
3. How can we call for privacy-preserving WebID analytics through types of semantics-enabled policy enforcement?
4. How to provide an effective and flexible service platform for data analysts through integrating R+SPARQL for graph-parallel analytics and MapReduce paradigm for data-parallel analytics?


Research Goals

1. We argue why we choose decentralized but not the centralized online social networking architecture for WebID analytics.
2. How can we provide privacy-preserving *batch* and *interactive* WebID analytics with effective and flexible services for data analysts?
3. How can we call for privacy-preserving WebID analytics through types of semantics-enabled policy enforcement?
4. How to provide an effective and flexible service platform for data analysts through integrating R+SPARQL for graph-parallel analytics and MapReduce paradigm for data-parallel analytics?


Research Goals

1. We argue why we choose decentralized but not the centralized online social networking architecture for WebID analytics.
2. How can we provide privacy-preserving *batch* and *interactive* WebID analytics with effective and flexible services for data analysts?
3. How can we call for privacy-preserving WebID analytics through types of semantics-enabled policy enforcement?
4. How to provide an effective and flexible service platform for data analysts through integrating R+SPARQL for graph-parallel analytics and MapReduce paradigm for data-parallel analytics?

Research Goals

1. We argue why we choose decentralized but not the centralized online social networking architecture for WebID analytics.
2. How can we provide privacy-preserving *batch* and *interactive* WebID analytics with effective and flexible services for data analysts?
3. How can we call for privacy-preserving WebID analytics through types of semantics-enabled policy enforcement?
4. How to provide an effective and flexible service platform for data analysts through integrating R+SPARQL for graph-parallel analytics and MapReduce paradigm for data-parallel analytics?


Contributions

1. Propose the concept of the semantic WebID analytics pipeline for automated data protection and analytics.
2. Three types of semantics-enabled policy for access control, data handling, and data releasing, are designed and enforced to enable the effective and flexible privacy-preserving WebID analytics.
3. Data analysts can flexibly choose SDC techniques on the large-scale privacy-aware Social Semantic Web.
4. We show how to effectively proceed anonymized WebIDs' collection and analysis but still ensure their utility.


Contributions

1. Propose the concept of the semantic WebID analytics pipeline for automated data protection and analytics.
2. Three types of semantics-enabled policy for access control, data handling, and data releasing, are designed and enforced to enable the effective and flexible privacy-preserving WebID analytics.
3. Data analysts can flexibly choose SDC techniques on the large-scale privacy-aware Social Semantic Web.
4. We show how to effectively proceed anonymized WebIDs' collection and analysis but still ensure their utility.


Contributions

1. Propose the concept of the semantic WebID analytics pipeline for automated data protection and analytics.
2. Three types of semantics-enabled policy for access control, data handling, and data releasing, are designed and enforced to enable the effective and flexible privacy-preserving WebID analytics.
3. Data analysts can flexibly choose SDC techniques on the large-scale privacy-aware Social Semantic Web.
4. We show how to effectively proceed anonymized WebIDs' collection and analysis but still ensure their utility.


Contributions

1. Propose the concept of the semantic WebID analytics pipeline for automated data protection and analytics.
2. Three types of semantics-enabled policy for access control, data handling, and data releasing, are designed and enforced to enable the effective and flexible privacy-preserving WebID analytics.
3. Data analysts can flexibly choose SDC techniques on the large-scale privacy-aware Social Semantic Web.
4. We show how to effectively proceed anonymized WebIDs' collection and analysis but still ensure their utility.


Part II

Background

Problems with Centralized Online Social Networks

- Information silos on one site is not usable in the others.
- A user is "stuck": migrating to another application is hard.
- Users cannot choose what Web applications do with their data.
- New application must acquire a critical mass of data from scratch.
- Do not allow users control over how their personal information is disseminated, which results in potential privacy problems.

–Decentralization: The Future of Online Social Networking

What are Decentralized Online Social Networks?

- Desire properties:
  - Decouple application from data
  - Give end-users control over their own data
  - Infrastructure (or platform) providers with IaaS or PaaS
  - Social network service developers with SaaS for users (or data owners)
  - Minimize the trust footprint for users to enforce their own data usage and control policies across applications.

–A World Wide Web Without Walls, Krohn, M., et al.


Privacy-Aware Social Semantic Web

- We need a new policy-oriented Social Semantic Web architecture.
- A profile management service that could be run in the browser or via a third-party website.
- Allow users to edit the attributes across multiple platforms and sites.

–A Standards-based, Open and Privacy-aware Social Web, W3C Incubator Group Report 6th Dec. 2010

Why Integrating R+SPARQL and MapReduce?

- MapReduce paradigm only works for lightweight data-parallel analytics.
- Distributed/parallel R for heavyweight graph-parallel data analytics.
- Graph-parallel analytics for discovering various social network's degree centralities, and data-parallel analytics for data anonymizing and output perturbation.
- Integrating R+SPARQL and MapReduce brings heavyweight graph-parallel analytics of R and lightweight data-parallel analytics of MapReduce.


How to use R+SPARQL and MapReduce for Data Analytics?

- We leverage the Semantic Web's open technologies for Social Semantic Web data representation and access.
- In future, merging centralized social network graph data in JSON with JSON-Linked Data (JSON-LD) for integrated analytics.
- JSON-LD are stored in the HDFS of cluster computers for *batch* and *interactive* data analytics in R.
- SPARQL queries and filters FOAF datasets in JSON-LD for R to enable heavyweight data analytics.


Semantics-enabled Policies

1. Semantics-enabled policies are composed of ontologies and queries, where ontologies describe the *concepts* of privacy-preserving data analytics services, and queries *enforce* the above privacy principles.
2. Semantics-enabled policies are correspond to query restriction, data manipulation/anonymization, and output perturbation:
   - Access Control Policy (ACP)
   - Data Handling Policy (DHP)
   - Data Releasing Policy (DRP)


Semantics-enabled Policies

1. Semantics-enabled policies are composed of ontologies and queries, where ontologies describe the *concepts* of privacy-preserving data analytics services, and queries *enforce* the above privacy principles.
2. Semantics-enabled policies are correspond to query restriction, data manipulation/anonymization, and output perturbation:
   - Access Control Policy (ACP)
   - Data Handling Policy (DHP)
   - Data Releasing Policy (DRP)


Part III

Privacy-Preserving WebID Analytics


The Semantic WebID Analytics Pipeline

- In a six-stage lifecycle:
  1. Acquisition and recording
  2. Extraction, cleaning, and semantic annotation
  3. Representation, integration, and aggregation
  4. Modeling and analysis
  5. Query processing and disclosure for analytics
  6. Interpretation


The WebID Analytics Pipeline (conti.)

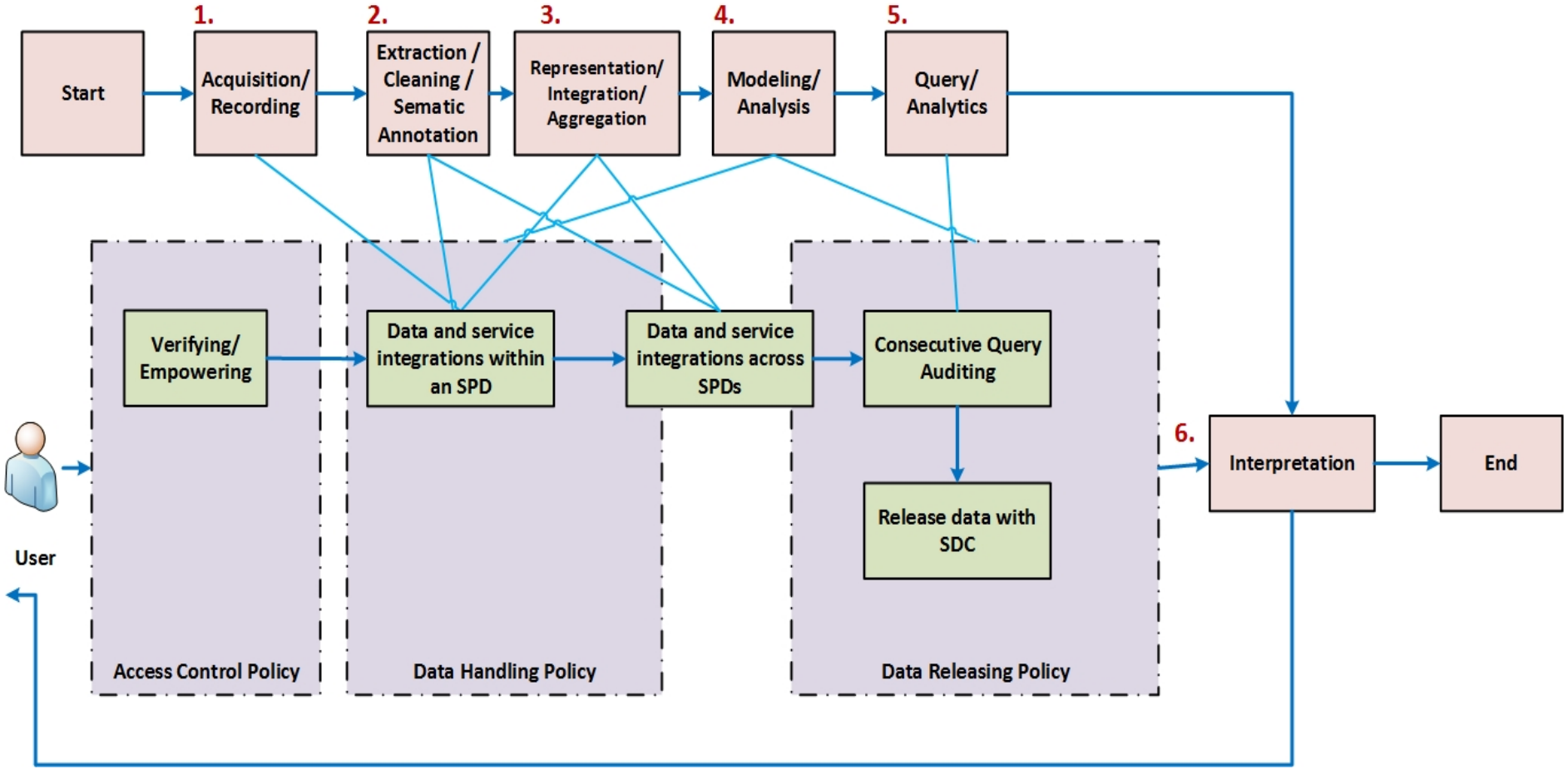


FIGURE: WebID Analytics Pipeline

Related Big Data Analytics Platforms

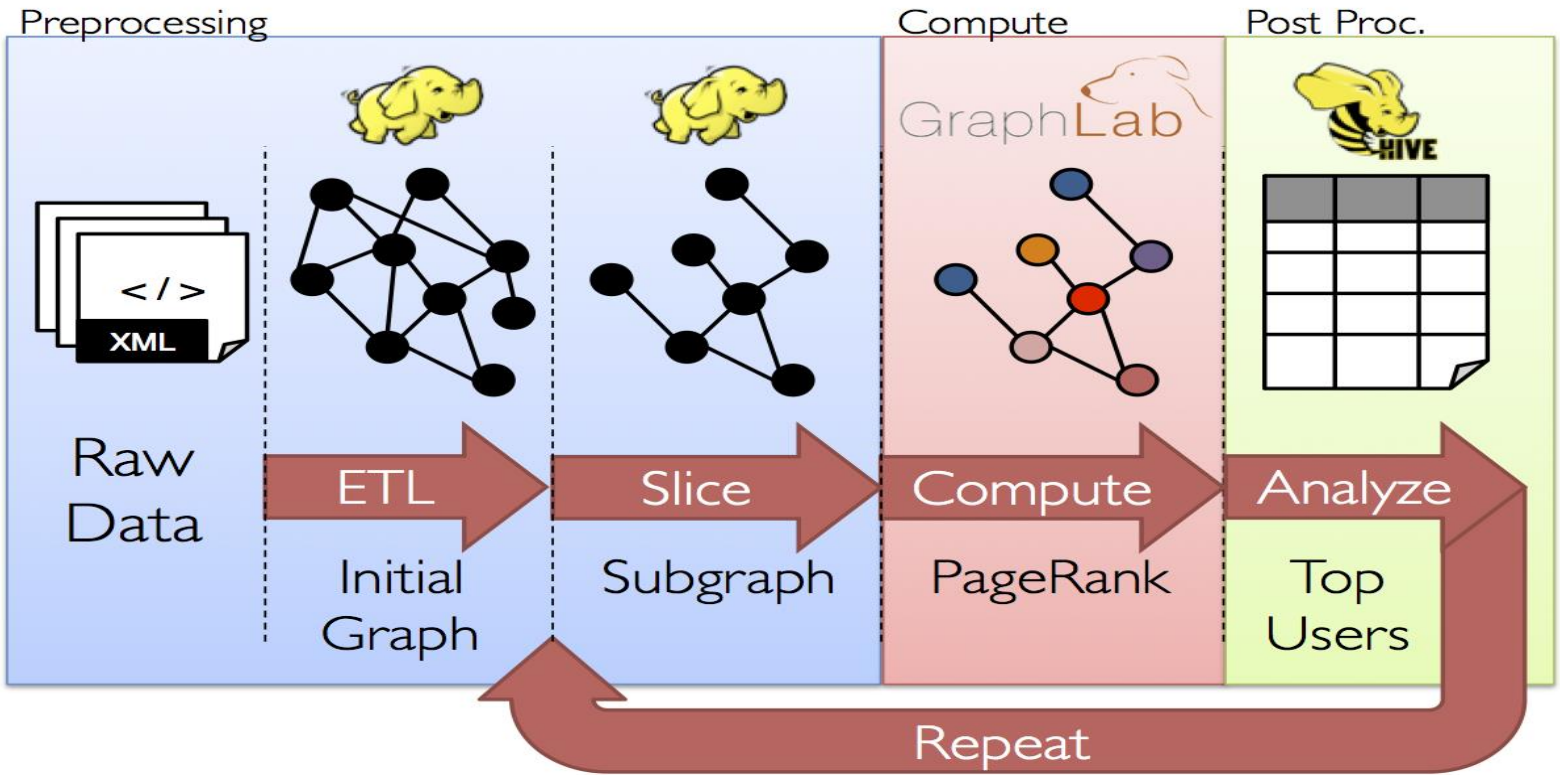


FIGURE: Berkeley AMP Lab. GraphX

Related Big Data Analytics Platforms

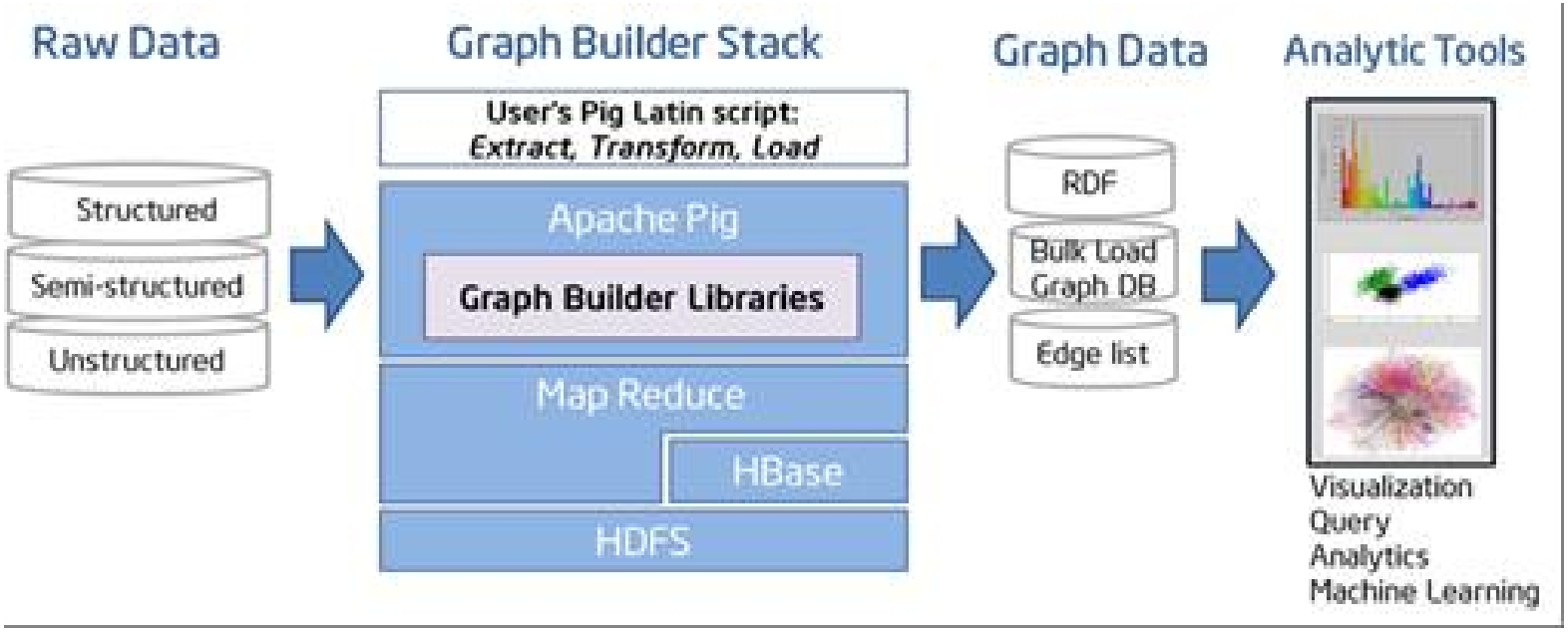


FIGURE:

Intel Lab GraphBuilder

Related Big Data Analytics Platforms

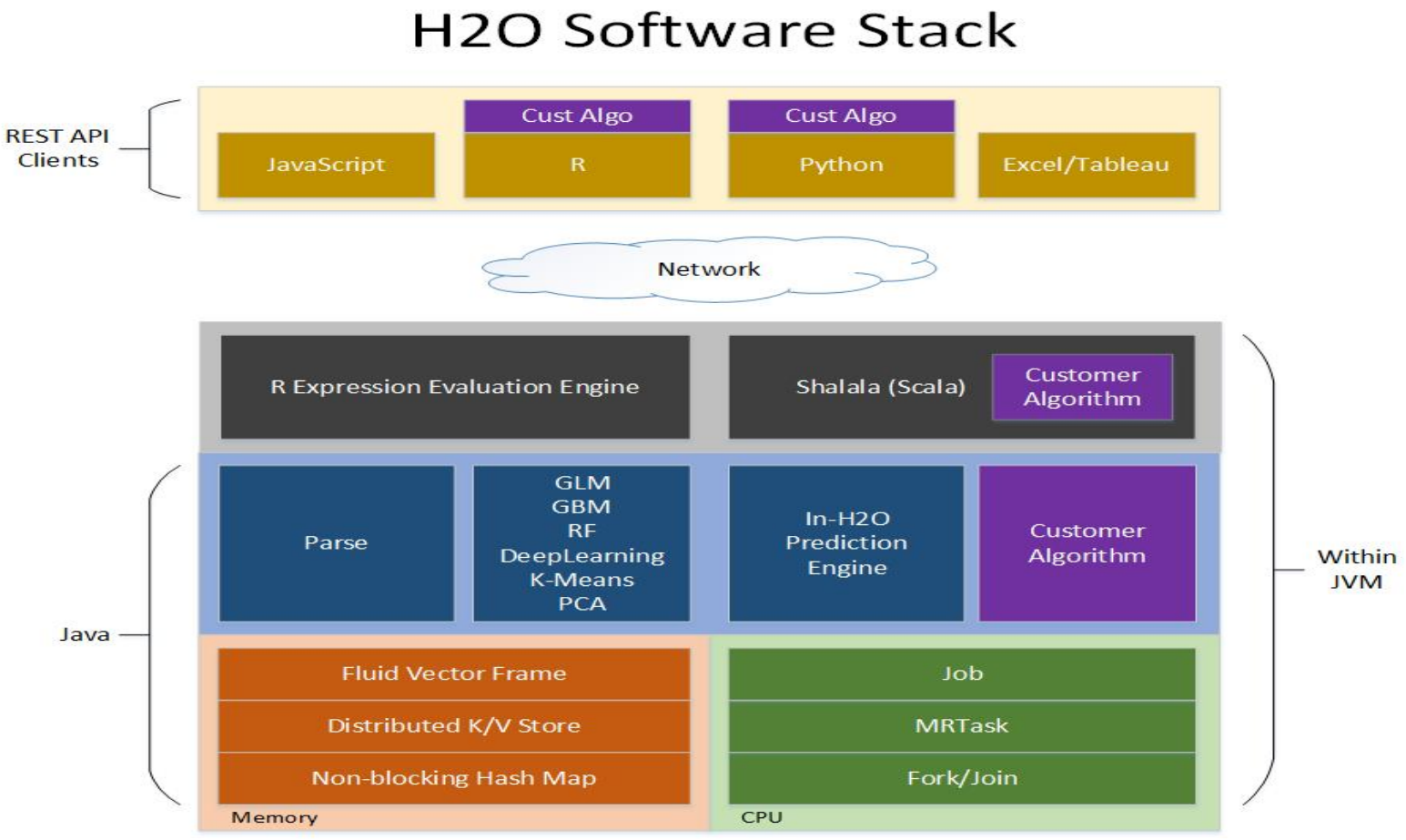
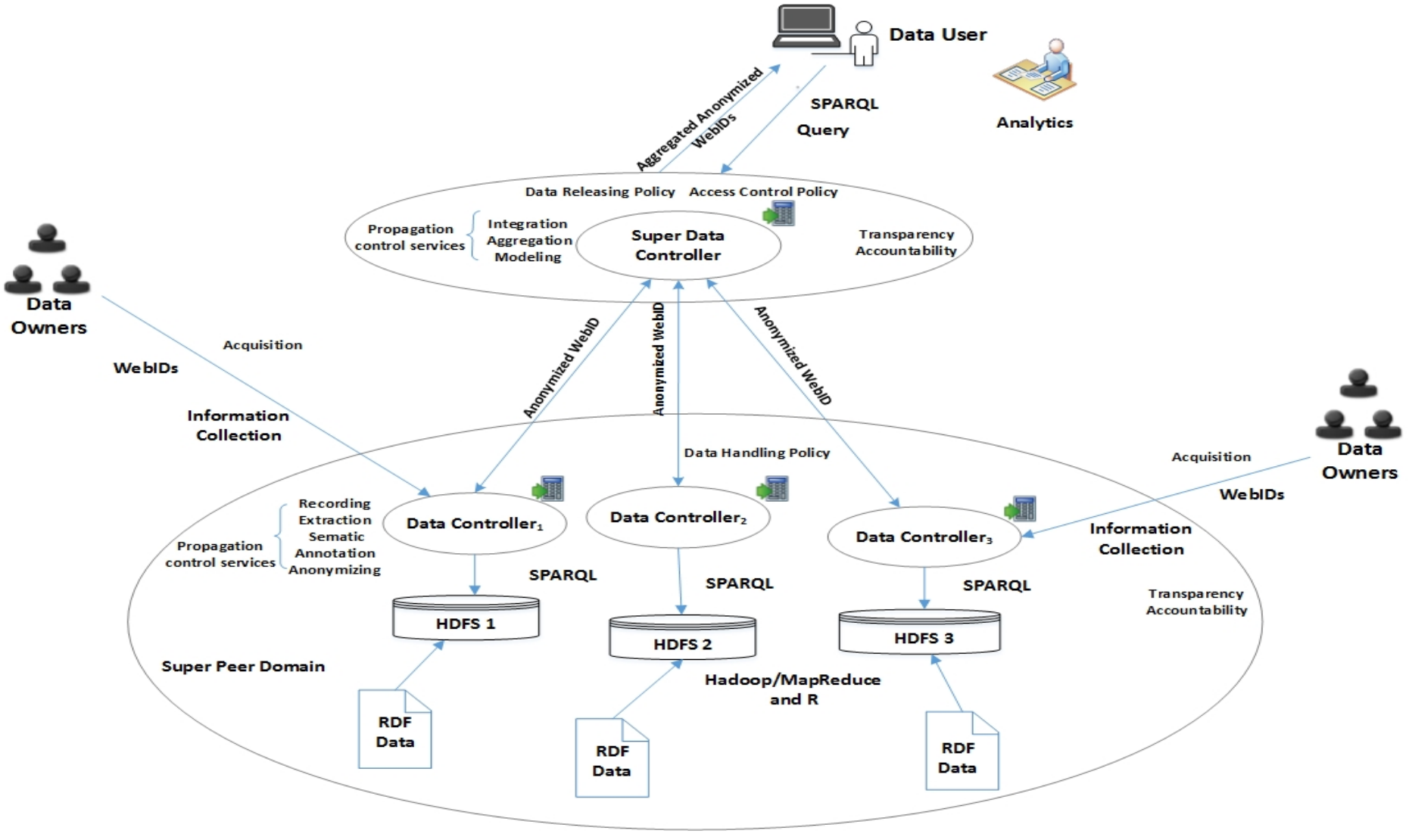


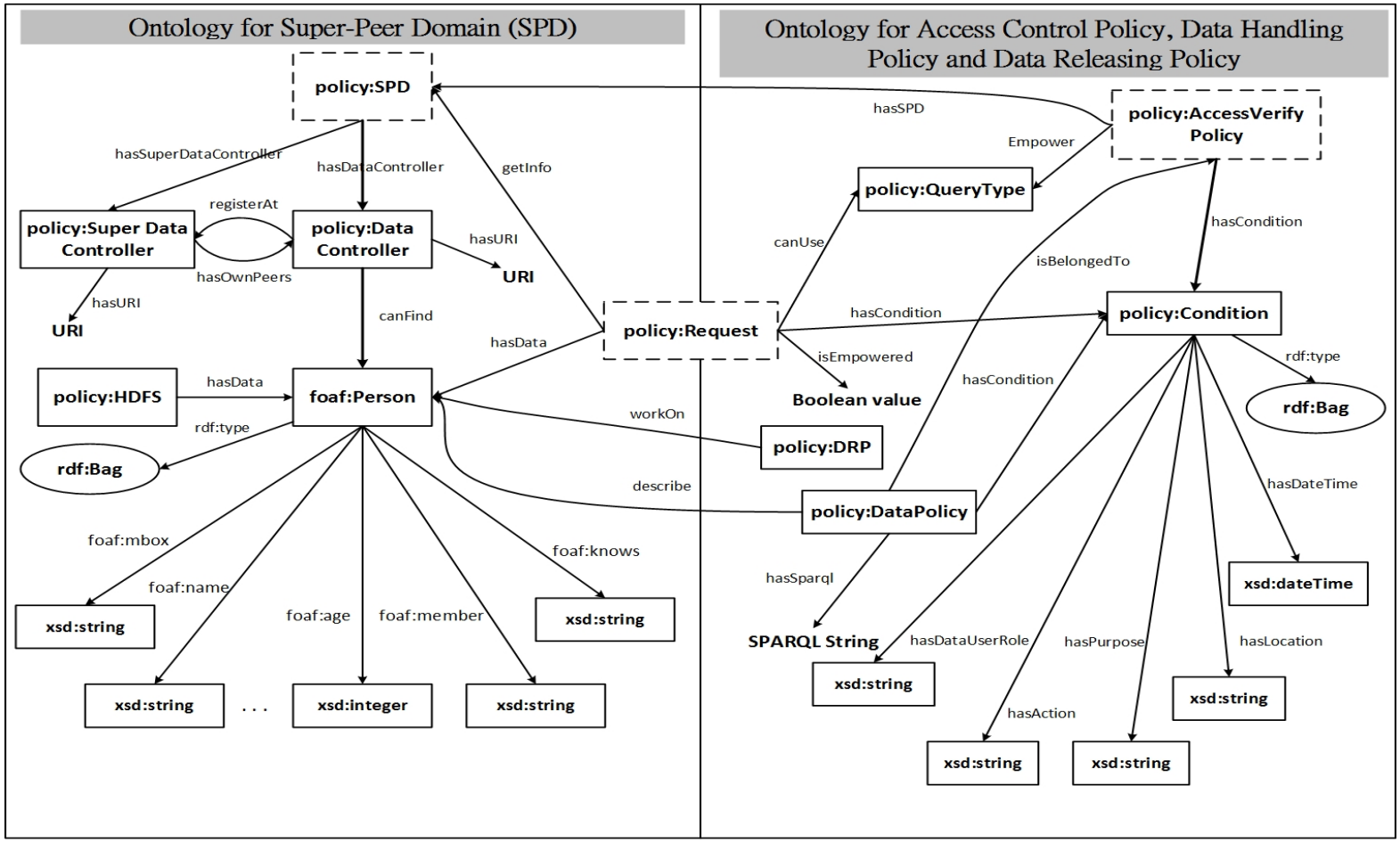
FIGURE:

H_2O Software Stack

A Super-Peer Domain (SPD) Data Cloud



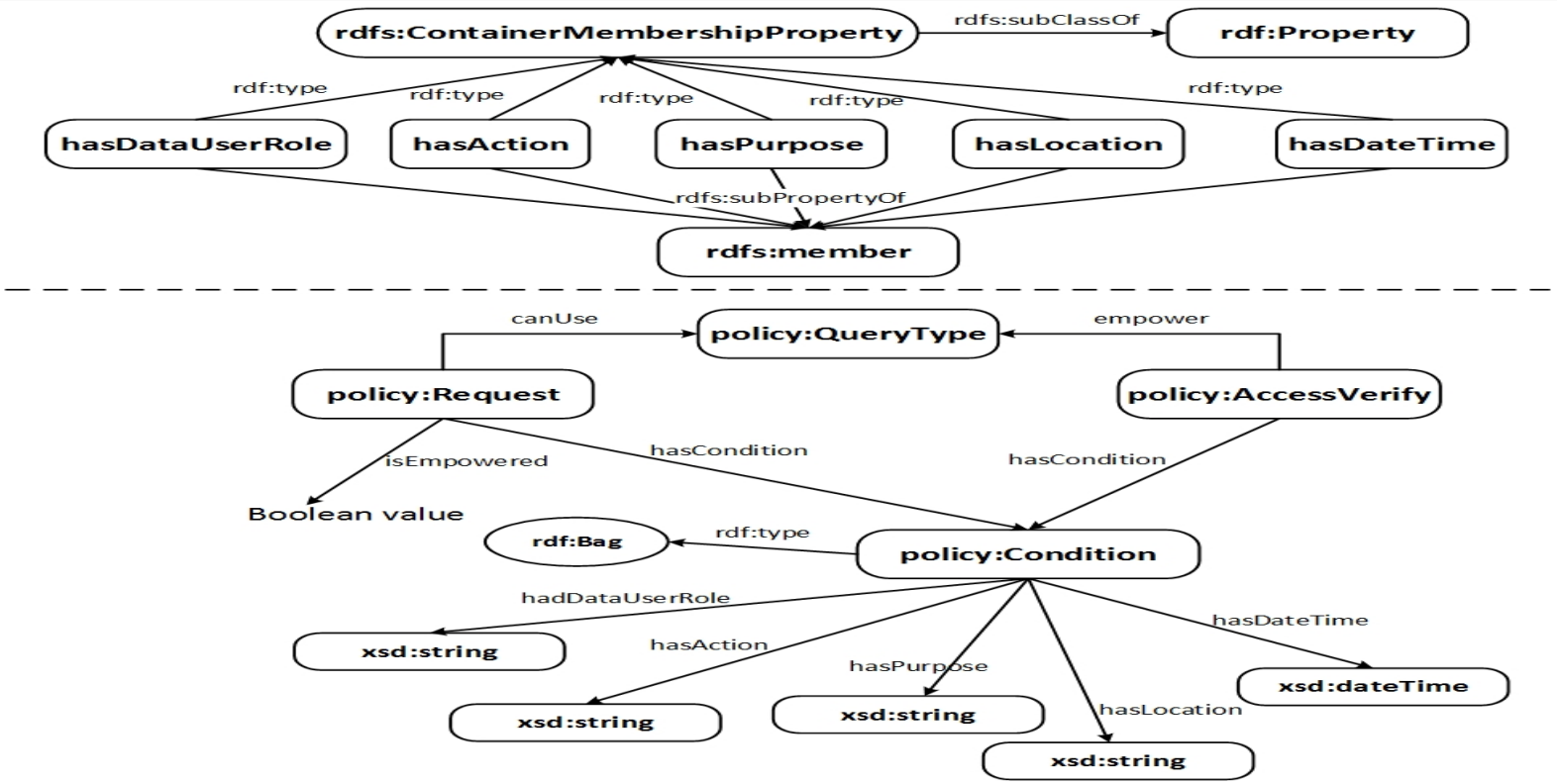
Policy Ontology for a Super-Peer Domain Cloud



An Ontology for Access Control Policy (ACP)

Definition of ACP Ontology

The concept of a data user's request verifications is represented as an ACP ontology and enforced as a SPARQL query.



A Query for Access Control Policy (ACP)

An ACP SPARQL Query

```
@prefix foaf : < http : //xmlns.com/foaf/0.1/ > .
@prefix policy : < http : //nccu.edu.tw/policy > .
policy : QueryType rdf : type rdf : Class
policy : PBQ rdf : type policy : QueryType
policy : Condition [
hasDataUserRole "DataAnalyst";
hasPurpose "Analytics";
hasAction "Read";
hasLocation "Taipei";
hasDateTime "2013 : 12 : 25 : 15 : 00" ].
```

A Query for Access Control Policy (ACP) (conti.)

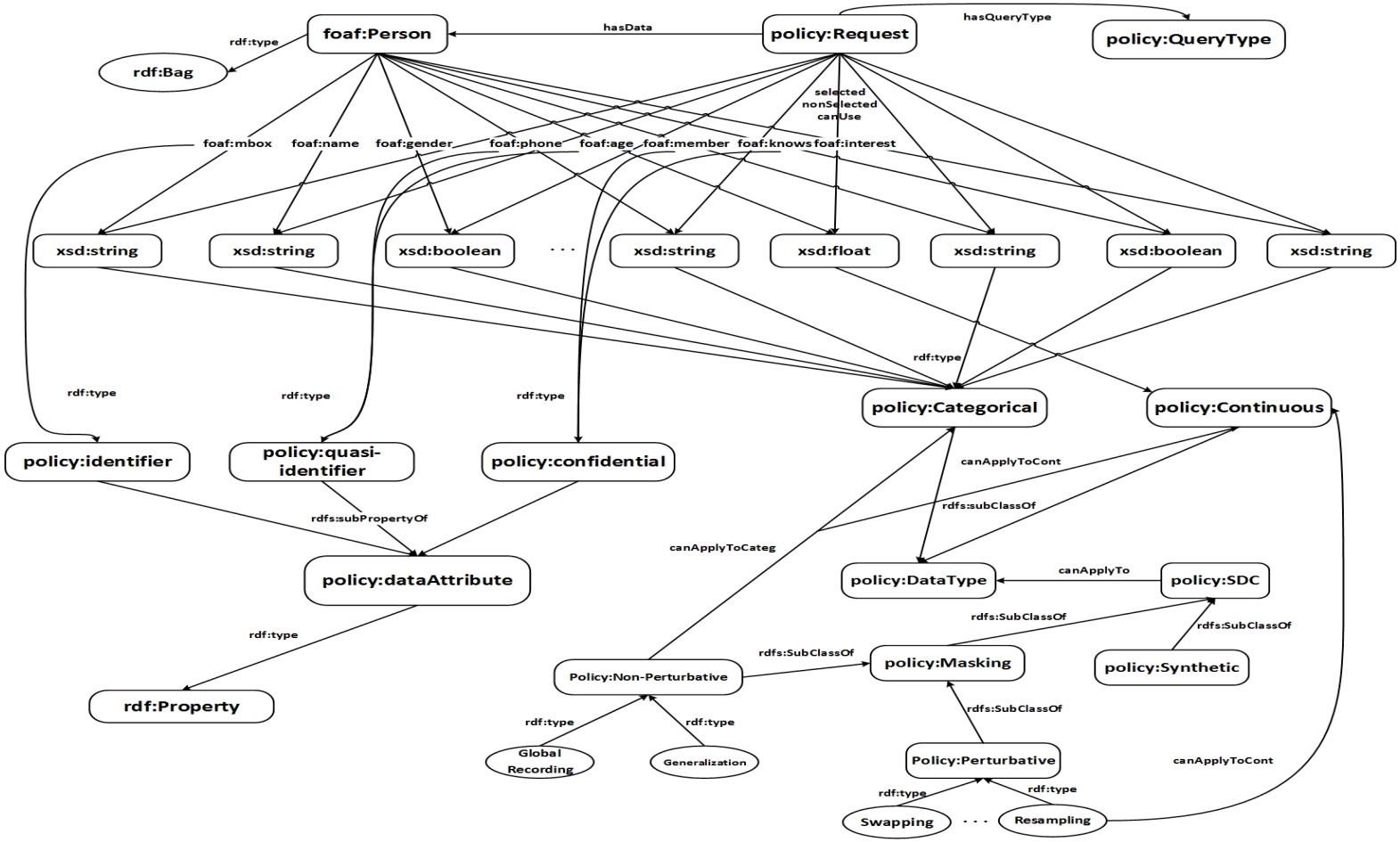
An ACP SPARQL Query (conti.)

```
Ask ?permit
From < PeterRequest.rdf >
Where {?r policy : isEmpowered ?permit.
?r [ ?qt rdf : type policy : QueryType;
policy : hasCondition ?c [
hasDataUserRole ?role;
hasPurpose ?purpose;
hasAction ?action;
hasLocation ?location;
hasDateTime ?time ] ].}
```

An Ontology for Data Handling Policy (DHP)

Definition of DHP Ontology

A DHP describes which SDC techniques are used to anonymize WebIDs' profile attributes, but not yet social network structure, in an FOAF property graph.



A Data Handling Policy (DHP)

A DHP calls for WebIds' profiles anonymizing

```
@prefix foaf : < http : //xmlns.com/foaf/0.1/ > .
@prefix policy : < http : //nccu.edu.tw/policy > .
< http : //nccu.edu.tw/j/foaf.rdf > a
foaf : PersonalProfileDocument.
< http : //nccu.edu.tw/j/foaf.rdf > foaf : maker :me.
< http : //nccu.edu.tw/j/foaf.rdf > foaf : primaryTopic :me.
:me a foaf : Person.
/ * De − identification * /
:me [ foaf : name "Yuh − Jong Hu";
foaf : homepage < http : //nccu.edu.tw/j >;
foaf : mbox < mailto : j@cs.nccu.edu.tw >;
/ * Generalization * /
foaf : phone < tel : +886 − 2 − 29387620 >;
```


A Data Handling Policy (DHP)(conti.)

A DHP calls for WebIds' profiles anonymizing

```
.....
foaf : knows [ a foaf : Person;
/ * De − identification * /
foaf : name "Kua − Ping Cheng";
rdfs : seeAlso
/ * enhanced microdata protection techniques * /
< http : //nccu.edu.tw/k/foaf.rdf > ].
foaf : knows [ a foaf : Person;
/ * De − identification * /
foaf : name "Ya − Ling Huang";
rdfs : seeAlso
/ * enhanced microdata protection techniques * /
< http : //nccu.edu.tw/y/foaf.rdf > ].
..... ]
```

An Ontology for Data Releasing Policy (DRP)

Definition of DRP Ontology

Governs acceptable conditions to disclose anonymized WebID's attributes to ensure the compliance of privacy protection principle.

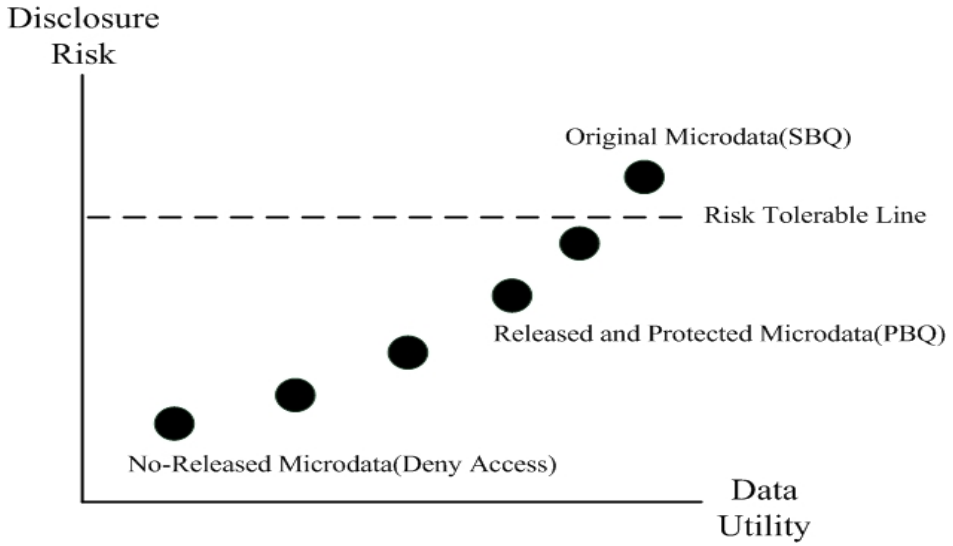
A DRP queries for anonymized WebID

```
Select ?graph ?gender ?age ?member ?interest
From < http : //nccu.edu.tw/j/foaf.rdf >
From named graph???
Where{< http : //nccu.edu.tw/j/foaf.rdf#me >
foaf : knows ?X.
{ ?X rdfs : seeAlso ?graph.
graph ?graph {[ a foaf : Person.
/ * De − identification * /
foaf : mbox ?mbox;
foaf : name ?name;
foaf : gender ?gender;
.....;
/ * Generalization * /
foaf : phone ?phone;
/ * GlobalRecording * /
foaf : age ?age;
foaf : member ?member;
foaf : interest ?interest;

foaf : knows [ ?graph ]. ]}}}
```

Semantic Data Protection Protection and Analytics

- Improve the situation, where SDC enforcement is obliged to original data providers and a data analytics user lacks the flexibility to choose suitable SDC methods.
- Seek a balance between a data owner's right for privacy protection and a data user's need for data analytics through transparency of SDC methods selection.
- Semantics-enabled DHP and DRP call for feasible SDC methods and ensures maximum data utility with a tolerable data disclosure risk.



Conclusion and Future Works

- Preliminary Results:
  1. Semantics-enabled policies are proposed and verified to provide query restriction, data manipulation/anonymization, and output perturbation.
  2. The R+SPARQL (for graph-parallel) and MapReduce (for data-parallel) platform is establishing to have a flexible and effective privacy-preserving WebID analytics.
  3. Billion Triples Challenge 2012 datasets for the FOAF/WebID analytics.
  4. A simple balance between data protection and utility through the semantics-enabled policies to call for suitable SDC methods.


Conclusion and Future Works

- Preliminary Results:
  1. Semantics-enabled policies are proposed and verified to provide query restriction, data manipulation/anonymization, and output perturbation.
  2. The R+SPARQL (for graph-parallel) and MapReduce (for data-parallel) platform is establishing to have a flexible and effective privacy-preserving WebID analytics.
  3. Billion Triples Challenge 2012 datasets for the FOAF/WebID analytics.
  4. A simple balance between data protection and utility through the semantics-enabled policies to call for suitable SDC methods.


Conclusion and Future Works

- Preliminary Results:
  1. Semantics-enabled policies are proposed and verified to provide query restriction, data manipulation/anonymization, and output perturbation.
  2. The R+SPARQL (for graph-parallel) and MapReduce (for data-parallel) platform is establishing to have a flexible and effective privacy-preserving WebID analytics.
  3. Billion Triples Challenge 2012 datasets for the FOAF/WebID analytics.
  4. A simple balance between data protection and utility through the semantics-enabled policies to call for suitable SDC methods.

Conclusion and Future Works

- Preliminary Results:
  1. Semantics-enabled policies are proposed and verified to provide query restriction, data manipulation/anonymization, and output perturbation.
  2. The R+SPARQL (for graph-parallel) and MapReduce (for data-parallel) platform is establishing to have a flexible and effective privacy-preserving WebID analytics.
  3. Billion Triples Challenge 2012 datasets for the FOAF/WebID analytics.
  4. A simple balance between data protection and utility through the semantics-enabled policies to call for suitable SDC methods.


Conclusion and Future Works (conti.)

- Future Works:
  1. Fully enforcing the semantics-enabled policies on the R+SPARQL and MapReduce paradigm platform.
  2. Including social network structure anonymizing to provide social link and identity anonymity.
  3. Using differential privacy technique for output disclosure control.
  4. Crafting an optimized balance between WebID protection and utility.