From sif to SOFA

Andrew Simpson
(and David Power, Douglas Russell and Mark Slaymaker)

Oxford University Computing Laboratory

June 18th, 2010

1 Motivation

2 sif

3 SOFA

4 Applications

5 Demonstrations

6 Summary and conclusions

Motivation

- Increasingly, there is a drive in many contexts to combine disparate data sets
- Often, the distribution of, and responsibility for, data reflects organisational structures
- Typically, there are issues of (systems, syntactic and semantic) heterogeneity to overcome
- And then there are issues pertaining to integration with legacy systems

Motivation

Our research has two broad goals:

- The facilitation of data aggregation across distributed, heterogeneous data sources
- The provision of secure, assured data sharing

sif

- sif (service-oriented interoperability framework): developed within the TSB-funded GIMI (Generic Infrastructure for Medical Informatics) project
- Based on experiences from e-DiaMoND and NeuroGrid
- Takes a data-agnostic approach
- Acts as a combined security and federation layer
- Facilitates the secure sharing and aggregation of data from (more or less) any structured data source
- Based on Java and web services

Value

- Low cost
- Limited impact
- Data ownership remains unchanged

Patterns of use

- ‘Secure’ piplines
- ‘Windows’ on research data
- Lightweight federation
- Integration of central systems with outliers

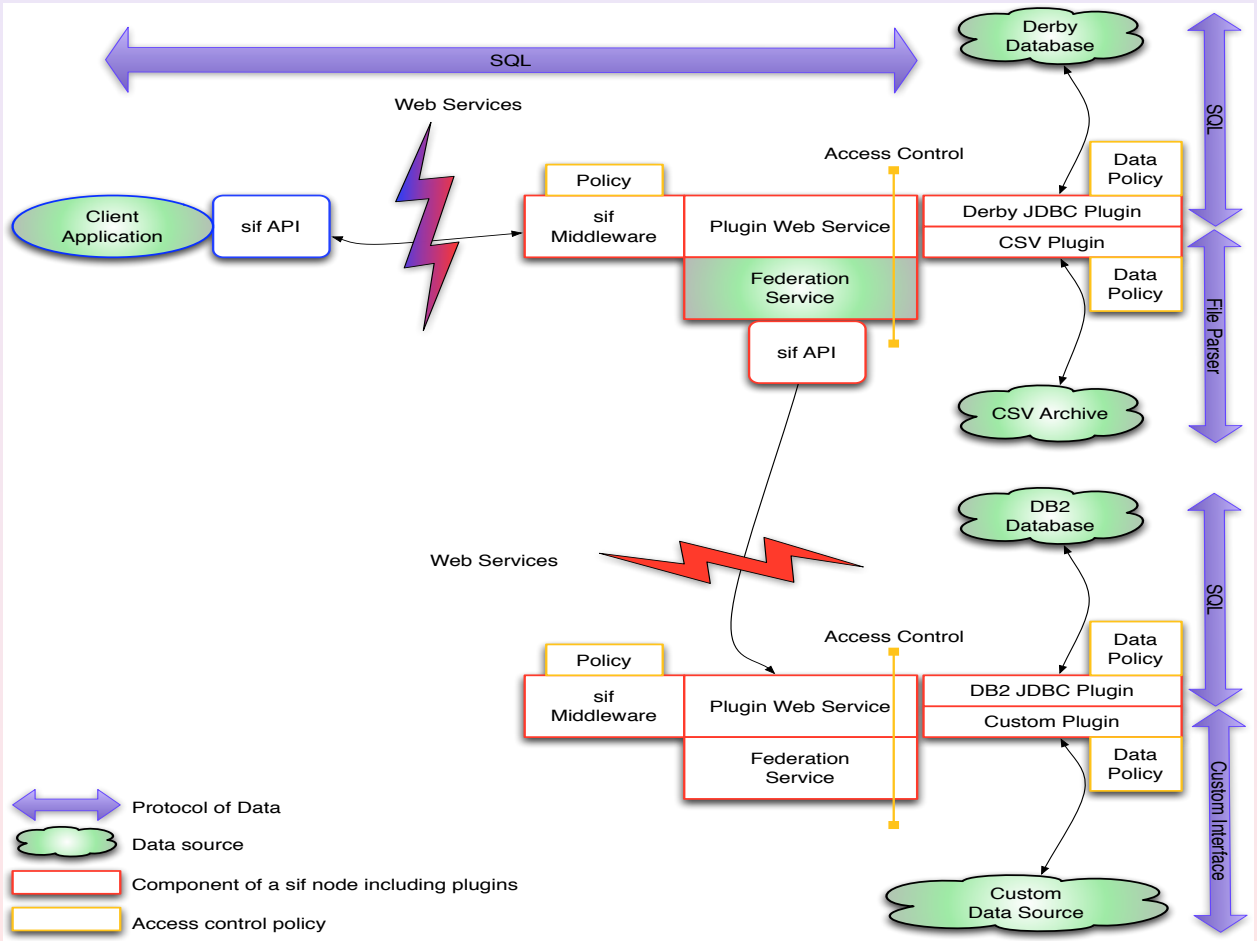
Plug-ins

- sif offers support for three types of plug-in: data plug-ins, file plug-ins and algorithm plug-ins
- By using a standard plug-in interface for each of the three types it becomes possible to add heterogeneous resources into a virtual organisation via sif
- There is no need for the resource being advertised through the plug-in system to directly represent the physical resource
- What is advertised as a single data source may come from any number of physical resources, or even another distributed system

The API

- The sif API allows applications to make calls to web services and receive results in a standard fashion
- For example, query results are returned as a WebRowSet, which can be used as normal ResultSet in Java, along with some additional information on the success (or otherwise) of the query (or, in the federated case, subqueries)
- The API also allows the query to be built up from objects, or be provided as an XML document
- The choice is provided to allow the application developer to choose the approach that they are most comfortable with

Architecture



Generic query tools

- A portal (very accessible; limited flexibility)
- A ‘query builder’ (more expressive; requires query-writing ability)

Access control

- sif allows the construction and enforcement of fine-grained access control policies
- XACML is leveraged
- **Advantage:** flexibility and expressiveness
- **Disadvantage:** verbosity and complexity

```
<Rule
RuleId="R_CanNotReadColumnC"
Effect="Deny">
      <Target>
            <Subjects>
                  <Subject>
                        <SubjectMatch
                        MatchId="urn:oasis:names:tc:
                        xacml:1.0:function:string-equal">
                              <AttributeValue
                              DataType="http://www.w3.org/
                              2001/XMLSchema#string">
                                    oX
                              </AttributeValue>
                              <SubjectAttributeDesignator
                              AttributeId="organisation"
                              DataType="http://www.w3.org/
                              2001/XMLSchema#string"/>
                        </SubjectMatch>
                  </Subject>
            </Subjects>
            <Resources>
                  .
                  .
                  .
```

SOFA

- Service Oriented Federated Authorization
- Runs for one year (from January 2010–December 2010)
- Funded by JISC
- Two key deliverables
  - ‘Lifting’ sif’s data-agnostic approach to the area of access control
  - Tool support for policy construction, analysis and transformation

Example

- Data source 1: college database; IBM DB2; fine-grained policies
- Data source 2: department database; MS SQL Server; role-based access control
- Data source 3: central administration database; MySQL; access control list

Tools and technologies for secure data sharing

- Exposing data:
  - Plug-ins for each data source
  - Access control policies, generated via the policy construction tool
- X.509 certificates associated with users
- Accessing data:
  - Query tool allows the construction of queries
  - Middleware facilitates federated queries and returns results
- Logical linking undertaken by end-user

Applications

- Student administration
- Heart modelling
- Research into Bipolar disorder
- Also:
  - *Data Security* case study

Demonstration 1: the policy tool

Demonstration 2: federated querying

Summary and conclusions

- sif: middleware framework that facilitates the secure sharing and aggregation of data from disparate, heterogeneous data stores
- SOFA: an extension of sif that allows data owners to leverage their access control paradigm of choice
- Applications: student administration; heart modelling; research into Bipolar disorder
- Immediate future work:
  - Refining the policy construction tool
  - Linking the tool with ongoing work pertaining to the formal analysis of access control policies
  - An initial open source release of SOFA