GIDS: Global Interlinked Data Store

Gareth Jones & Dave Braines
Emerging Technology Services
IBM United Kingdom Ltd
Hursley Park, Winchester, UK
Email: {garethj,dave_braines}@uk.ibm.com

Paul Smart & Trung Dong Huynh
School of Electronics & Computer Science
University of Southampton
Southampton, UK
Email: {ps02v,tdh}@ecs.soton.ac.uk

Jie Bao
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY, USA
Email: baojie@cs.rpi.edu

***Abstract*— This paper introduces the Global Interlinked Data Store[1] (GIDS), a technique to support the easy creation and retrieval of interlinked semantic data within a web-scale distributed network environment such as the World Wide Web (WWW). The GIDS enables the network to be treated as a data store without worrying about files, databases or other traditional data storage concerns. Data created on the network can be subsequently accessed and navigated by end users and software agents alike. The GIDS proposes a novel three-stage data storage process which enables the data to be stored in up to three contextually relevant locations to enhance subsequent retrieval opportunities. We believe that the capability offered by the GIDS will be of significant use to rapidly formed diverse coalitions who wish to communicate and exchange semantic data in a large network environment such as the WWW.**

I. Introduction

The GIDS proposes a technique for directly storing semantic data "on the surface" of a network and enables the subsequent navigation of that data. This is achieved through the use of network resolvable addresses (dereferenceable URIs) in the definition of the data, removing the need for separate indexes defining the location of these data, and removing concerns about specific file formats, file locations or access to triple stores.

The GIDS also enables data to be spread across referenced locations allowing a contextually relevant distribution of data across the network, and provides a basis for multiple paths of access to the data for subsequent onward navigation. The overall aim for the GIDS is to ensure that every entity (and every triple) is directly accessible via the corresponding dereferenceable URI meaning that a simple network request to that URI will yield a response containing triples which in turn contain dereferenceable URIs, the details of which can be requested through subsequent network requests in a recursive manner.

The GIDS is intended to be a supporting platform for the Semantic Web, providing a simple set of interfaces within which data can be written, read and navigated by software agents, applications, or human users. The GIDS is therefore primarily concerned with RDF data storage and does not in itself provide any of the higher level capabilities associated with the Semantic Web (such as inference/entailment, complex queries, etc).

The GIDS proposes three important capabilities which build upon the basis of the Semantic and Linked Data Web, specifically:

A defined response format: When dereferencing a URI the format and structure of the expected response is known and takes the form of triples which are related to the entity that corresponds to the requested URI.

Direct triple access interface: The ability to directly read and write semantic data triples to/from the network through simple network resource requests (e.g., a HTTP GET request to a particular URL in the same way that a browser requests a web page).

Multiple storage locations: The GIDS proposes the ability to store semantic data triples in up to three logical locations on the network. Each of the dereferenceable URIs for the three components in the triple can be notified of the triple when it is asserted and can then optionally store the details of that triple for subsequent processing.

The combination of the above three core capabilities yields this new approach which enables data to be easily stored *directly on the network* through simple network requests, to be retrieved in the same way with *a known response format* to facilitate subsequent automated processing, and enables each entity referenced by a triple to *store a separate copy of that triple* to facilitate improved navigation.

II. Example

Figure 1 shows a simple example of the GIDS being used to store data about a number of entities across nodes within a distributed network environment. A node is discriminated based on the hostname component of each URI. Node 1 defines and stores information related to three entities (A, B, C) and Node 2 defines and stores information related to three entities (D, E, F). The information is stored in the form of triples with s, p, and o being the subject, predicate and object respectively. A triple of particular interest *(A, B, D)* is actually stored against two separate nodes (Node1 and Node2). This is because Node1 defines entities A and B, whereas Node2 defines entity D. Using the GIDS each entity involved in a triple will be notified when the triple is asserted, and in this example both nodes therefore have stored a copy of the triple. The diagram also shows other nodes and other entities,

[1]This is the short paper, however the original full size paper will be published after the ACITA'09 conference as technical report with the same title. Please refer to the technical report for a full exploration of the underlying principles, details of military relevance and a description of the prototype implementation.

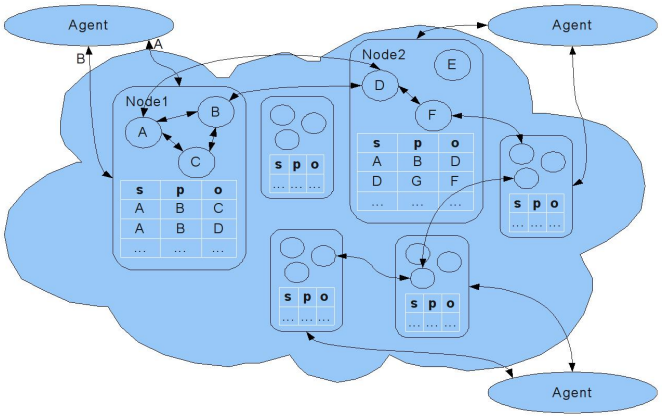

Figure 1. An example of triples distributed using the GIDS

and indicates various agents (software or human) who are interacting with the entities and triples via standard GIDS network requests.

III. Reification

The GIDS also provides capability for the triples themselves to be referred to, providing the basis for an important capability known as "reification". Upon creation each triple is assigned a unique identifier in the form of a dereferenceable URI and this identifier is then returned in all future query responses, allowing subsequent access to the triple. This reification capability enables statements to be made about the original triple statement.

Using this simple reification technique enables unlimited chains of reified information to be stated and the GIDS reification approach is designed to closely match the de-facto *rdf:Statement* reification within RDF.

IV. Advantages, Disadvantages and Assumptions

The GIDS introduces a number of distinct advantages above current Semantic Web and Linked Data Web approaches:

Data Visibility: All data can be accessed, modified and deleted through the interaction with URIs. This enables easier data sharing, re-use and integration and is a significant extension to the current Linked Data Web approach.

Data Access: The availability of data from all entities associated with a data element (a triple) allows data access from any of these entities. This provides entry points to users of the data from either the subject, predicate, or object entity.

Data Ownership: Data in the GIDS is stored and owned by the entities associated with the data. This approach enables automated capture of predicate related information to occur in addition to the more localised ownership of data relating to the subject and object entities.

Data Redundancy: In GIDS, a triple can be replicated at the subject, predicate and object entities thus providing some data redundancy. If data is lost, was never stored, or is temporarily unavailable at one of the associated entities it can be retrieved from other entities associated with the required triple.

Some necessary disadvantages are also introduced:

Efficiency: Every request to create or delete a triple can be made to all three URIs that make up that triple, i.e. to the subject, predicate and object URIs. This means that every logical request may result in up to three physical requests across the network.

Reliability and trust: Since we are operating in a WWW environment, servers may become unavailable or resource-constrained, they may be unreliable, data may be inconsistent between different sources or may be untrustworthy, etc.

Privacy: The notification of data to each of the three associated URIs does raise concerns about privacy, since it may not always be appropriate to notify a URI that a related statement has been made. The GIDS does not enforce the notification to all associated URIs, so the application or user can decide which of the three to send. Also, the GIDS does not mandate that a URI which receives a request to assert a triple must act on that request.

The GIDS makes some assumptions; some are key enablers for the desired capabilities, whilst others are inherited from the Semantic and Linked Data Web context:

Dereferenceable URIs: As with the Linked Data Web a restriction is imposed on the use of dereferenceable URIs for all entities.

Anonymous/Blank Nodes: The GIDS does not readily support blank or anonymous nodes since they do not have a globally guaranteed unique identifier to act as the dereferenceable URI. Since the GIDS is aimed at the storage of triple data we delegate the problem of dealing with blank nodes to the application or agent using the GIDS.

Inference/Entailment Support: The GIDS does not offer any inferential or entailment capability since it is concerned purely with the storage and retrieval of triple data in a distributed network environment and again this behaviour is delegated to the application or agent using the GIDS.

V. Implementation

There is a prototype implementation of the GIDS framework which is based upon a RESTful implementation, but this is not described in this short version of the paper.

VI. Conclusion

The Global Interlinked Data Store (GIDS) provides a novel approach to support the creation and navigation of semantic data triples in a distributed network environment. The GIDS provides a number of extensions to the current Linked Data Web approach, and is designed to provide a lightweight and simple interface to semantic data stored in a network environment such as the World Wide Web.

Acknowledgment

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.