# SIOC: an approach to connect web-based communities

John G. Breslin*, Stefan Decker, Andreas Harth and Uldis Bojars

Digital Enterprise Research Institute
National University of Ireland, Galway
University Road, Galway, Ireland
Fax: +353 91 512541
E-mail: john.breslin@deri.org
E-mail: stefan.decker@deri.org
E-mail: andreas.harth@deri.org
E-mail: uldis.bojars @deri.org
*Corresponding author

**Abstract:** Online communities are islands of people and topics that are not interlinked. Complementary discussions exist on disparate systems but it is currently difficult to exploit the available distributed information. A Semantically Interlinked Online Community (SIOC) can enable efficient information dissemination across communities by creating an ontology that will model concepts identified in discussion methods. Data instances can be accessed from community sites using this ontology, enabling connections between local and remote concept instances, and allowing queries on, or transfer of, the data. By searching on one forum, the ontology and interface will allow users to find information on other forums that use a SIOC-based system architecture. Other uses include cross-site querying, topic-related searches, and the importing of SIOC data into other systems. Fusing information and inferring links among various applications and types of information with SIOC provide relevant insights that make the community information available on the internet more valuable.

**Keywords:** weblogs; semantic web; ontologies; forums.

**Reference** to this paper should be made as follows: Breslin, J.G., Decker, S., Harth, A. and Bojars, U. (2006) 'SIOC: an approach to connect web-based communities', *Int. J. Web Based Communities*, Vol. 2, No. 2, pp.133–142.

**Biographical notes:** John G. Breslin received his PhD from the National University of Ireland, Galway (NUI Galway). He is a Postdoctoral Researcher at the Digital Enterprise Research Institute (DERI), NUI Galway. His research interests include social networks and online communities.

Stefan Decker received his PhD from the University of Karlsruhe, Germany. He is a Research Fellow, Adjunct Lecturer at DERI, NUI Galway. His research interests include the semantic web and P2P technologies.

Andreas Harth is currently studying for his PhD at DERI, NUI Galway. His research interests include RDF storage and querying.

Uldis Bojars is also studying at DERI, NUI Galway. His research interests include semantic matching of skills and community discussions.

## 1 Introduction

Computer-supported collaboration and discussion systems for closed and open domains are in widespread use on intranets and the internet. The closed community collaboration model usually has a limited and controlled audience where restricted information access and workflow management are the main requirements (for example, commercial groupware products, such as Lotus Notes for businesses, or open source CSCW (Grudin, 1994) products, such as NetOfRce for researchers). The open community collaboration model facilitates information exchange with emphasis on open involvement, participation, circulation of information and feedback (examples include public bulletin boards or archived mailing lists, Usenet newsgroups, social networks, weblogs and wikis).

Online communities (McArthur and Bruza, 2001) using open collaboration systems have the potential to replace the traditional means of keeping a community informed via libraries and publishing (Millen, 2000). These sites allow improved communication and interactive contact within a community, by providing an online collaboration space for members to find and contribute certain interest-related or regional information (Wellmann and Gulia, 1999).

However, it is difficult to exploit the available information in such community sites on the internet, especially when most online communities are hosted on stand-alone sites that cannot be interconnected due to application and interface differences. Also, each community site will normally have a unique entry point to its own discussions. Parallel discussions on interrelated topics may exist on a number of sites that are not linked. There is a huge amount of related information that could be harnessed across such online communities, from similar member profile details to common-topic discussion forums.

This paper addresses how to maximise the usage of this potentially valuable information and how to enable the location of relevant information in online communities. The research question is to identify and model the concepts found in online discussion methods, and to create a data infrastructure among different community sites. This will aid in a reduction of the information overload from existing search engines, and will use semantic web technologies to make the information useable by applications.

SIOC faces the following interesting challenges:

- The grand challenge is adoption by community sites, *i.e.*, how a critical mass can be reached by enticing users to make use of the SIOC ontology. By using concepts that can be easily understood by site administrators, and by providing properties that are automatically created by an end user, the SIOC ontology can be adopted in a useful way.
- A second challenge is how best to use SIOC with existing ontologies and collaboration technologies. This challenge can be partially solved by mappings and interfaces to commonly used ontologies, such as DC[1] FOAF[2] and RSS,[3] and by wrappers to technologies such as NNTP, SMTP and SQL databases.

SIOC also has to deal with the addition and removal of topics. An evolving category hierarchy is essential for any living online community, but this presents challenges with respect to the matching of recently created or deprecated topics.

Another challenge is how well SIOC will scale. If there are more sites to query, then there are more potential relevant results, but also longer response times and higher load on the participating community sites. We must keep this scaling challenge in mind when creating the architecture of an interconnected system of the community sites.

## 2 Why is this problem significant?

Some of the main problems in relation to existing online community technologies are:

- Information on community sites cannot be harnessed correctly by search engines (Bruza *et al*., 2000) limited to syntactic matching, *e.g*., by keyword matching on bulletin boards.
- Many isolated communities that discuss complementary topics exist.
- Information is being repeatedly requested across separate sites, and people are wasting time searching for relevant community information by waiting for answers that have already been posted elsewhere.

Linking individual posts with others is possible on the HTML level, but a forum search will not represent this link. Using the example in Figure 1, a user is searching for information on installing broadband on a Linux-based PC in their house in Galway. There is a post A discussing local ISPs on site 1, a bulletin board dedicated to Galway, that references (on the HTML level) both a Usenet post B comparing broadband modems and a mailing list post C detailing how to install broadband on Linux. Previously, the user would have had to traverse three sites to find the relevant information. However, by making use of the SIOC ontology and remote RDF querying, a search for broadband on the Galway bulletin board will also yield the relevant text from the interlinked Usenet and mailing list posts B and C.

In reality, explicit references between posts do not often exist, and therefore similarities among community sites could be quantified (for example in Figure 2, an explicit linking of sites could be weighted as 0.1, forums as 0.01, sites as 0.001, and the weighted combination is then used to determine sites with a ranking greater than 0.1 for a cross-community search).

Once there exist enough sites that have richer query facilities to instances of SIOC data, then these different sites can be interlinked. If a user has an account at site 1, then site 2 could pull that user information from site 1 and that user would not need to maintain his/her own accounts database. Other benefits include: applying reasoning facilities (such as those provided by OWL-S time (Pan and Hobbs, 2004)) to make use of events descriptions; visualisation of scheduling information of individuals or groups of people; and representations of where people related to a certain topic are located geographically.

**Figure 1** Retrieving related posts through SIOC

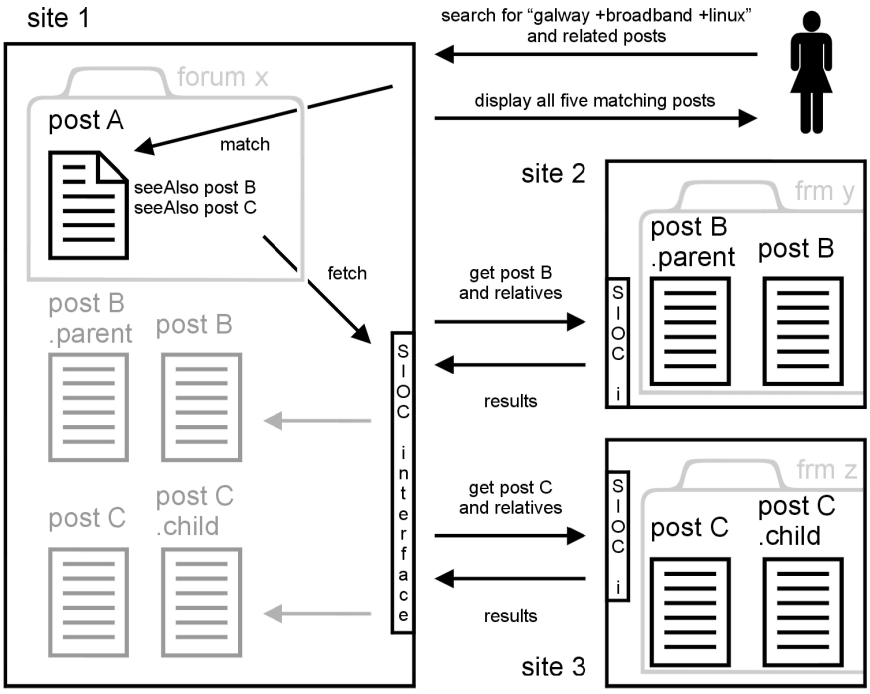


**Figure 2** Inferring implicit inter-site connections