Data Portability with SIOC and FOAF

Uldis Bojārs[1], Alexandre Passant[2,3], John G. Breslin[1]

[1] DERI, National University of Ireland, Galway, Ireland
firstname.lastname@deri.org
[2] LaLIC, Université Paris-Sorbonne, Paris, France
alexandre.passant@paris4.sorbonne.fr
[3] Electricité de France R&D, Clamart, France
alexandre.passant@edf.fr

Abstract. Data portability has become an important requirement on the Social Web. This paper describes how data and social network portability can be achieved by building upon existing Semantic Web developments such as SIOC and FOAF. The SIOC project provides a common format for expressing user-generated content in a machine-readable form and tools for import/export of this information, while the FOAF vocabulary provides a format for interlinking users' social networks. We demonstrate this approach on an example of a SIOC import plugin for WordPress.

Key words: Social Media, Semantic Web, Web 2.0, Data Portability, Social Networks, FOAF, SIOC

1 Introduction

Social media sites, including social networking services, have captured the attention of millions of users as well as billions of dollars in investment and acquisition. To better enable a user's access to multiple sites, portability between social media sites is required in terms of (1) identification, personal profiles and friend networks and (2) user's content expressed on each site, whether it is about blog posts, pictures, bookmarks or any type of data. Such portability would allow users to easily exchange content between services, or merge and share their social network between various websites.

Two Semantic Web projects SIOC (Semantically-Interlinked Online Communities) and FOAF (Friend of a Friend) can be combined to enable data portability between social media sites:

- The FOAF vocabulary allows us to represent people and their social networks, providing the social network component of data portability. It can be used in a combination with the the OpenID identity system.
- The SIOC vocabulary is an open format for expressing information about user-generated content in an interoperable way. It provides the content object component of data portability. The SIOC Types module can be used to further specify different types of Social Web / Web2.0 objects that we may want to describe [3].

In this paper, we will discuss the application of these technologies to address issues with portability between such services. We will then describe how this information can be ported to other sites with tools such as the WordPress SIOC Import plugin.

2 Overview of Data Portability

"Social network portability" is the term used to describe the ability to reuse one's own profile across various social networking sites. Brad Fitzpatrick[1] spoke from a developer's point of view about forming a *"decentralised social graph"* [7] and discussed some ideas for social network portability and aggregating one's friends across sites. As such, the social network portability becomes the first requirement for data portability. However, it is not just friends that may need to be ported across social media sites, but identity and user-created content items as well.

The DataPortability initiative[2] was launched in 2007, with members from various organisations including Facebook, Google and Microsoft coming together to discuss portability issues from technical and legal standpoints. It promotes the idea that individuals have control over their data by determing how they can use it and who can use it. This includes access to data that is under the control of another entity. Users should have the right to share their content items with other services and to move this content to other services if needed. We will refer to these items as one's social media contributions, or SMCs. Through such portability, the interactions and actions of a person with other users and objects (on systems they are already using) can be used to create new person or content associations when they register for a new social media site.

In the next section, we will describe how FOAF can be used to define user profiles and can act as a unique entry point for personal data across different social media sites. The second part of this paper will describe the data portability aspect, using the SIOC vocabulary as an open format to describe user-created content for moving data from one platform to another.

3 Social Network Representation with FOAF

3.1 Identity and Networking Management Across Social Media Sites

While many social media sites allow people to define their social networks, only a few of them permit users to export their networks so that they can be reused across other applications. Moreover, when this is the case, users have to rely on some specific APIs, which means writing ad-hoc tools for each data provider.

The FOAF project provides a way to represent social network data in a shared and machine-readable way, since it defines an ontology for representing people

[1] Founder of the LiveJournal blogging community
[2] http://dataportability.org

and the relationships that they share. Some sites already offer FOAF export, such as LiveJournal[3] and MyBlogLog[4]. For other sites, independent developers have created different tools (i.e. wrappers) to enable users get their data in a machine-readable for. For example, user profile information is available in RDF thanks to exporters for Flickr[5], Facebook[6] or Twitter[7].

Using FOAF, people and relationships can be modeled using these principles:

- each person is represented as a `foaf:Person` instance and may be assigned URI(s), their unique identifiers on the (Semantic) Web;
- each person has various properties, such as a name (`foaf:name`), nickname (`foaf:nickname`) or birthdate (`foaf:birthday`);
- people can be related to each other using the `foaf:knows` property.

For example, the following snippet of code represents one of the author's profiles created from Flickr using FOAF:

```
flickr:33669349@N00 a foaf:Person ;
  foaf:name "Alexandre Passant" ;
  foaf:mbox_sha1sum "528b95cc44060ceea571d7498a9fd2c7e3ca8a4c" .
  foaf:knows flickr:32233977@N00 .
```

Leveraging Semantic Web representations of people and social networks using widely-adopted ontologies such as FOAF allows us to use generic RDF parsers and SPARQL[8] (a RDF query language which recently became a W3C recommendation) to browse and reuse data.

Thus, end users can use the same tools to parse their network wherever it comes from. To that extent, FOAF simplifies the process of writing tools for developers of social-networking frameworks, especially since many open-source tools are available for most platforms[8] [9].

3.2 Merging and Querying Social Networks

As introduced previously, FOAF allows us to describe personal profiles and also to represent relationships between people. But there remains a problem that various sites may export a FOAF representation of users and social networks using their own URIs schemas, which means that there is still a need to merge and consolidate distributed profiles. In order to consolidate URIs, network owners may rely on the following properties to represent the identity of existing objects [2] (1) `owl:sameAs` is used to identify that two resources are the same in spite of different URIs and (2) `rdfs:seeAlso` is used to let crawlers and Semantic Web

[3] http://livejournal.com
[4] http://mybloglogb.typepad.com/my_weblog/2008/04/mybloglog-bring.html
[5] http://apassant.net/blog/2007/12/18/rdf-export-of-flickr-profiles
[6] http://www.dcs.shef.ac.uk/ mrowe/foafgenerator.html
[7] http://sioc-project.org/node/262
[8] http://www.mkbergman.com/?page_id=346
[9] http://www.w3.org/2001/sw/SW-FAQ#tools

browsers such as Tabulator [1] know where to find additional information about the resource. Many Semantic Web tools also follow Linked Data guidelines[10] and try to dereference instance URIs, thus providing another way to find additional RDF data.

Using these properties in a distributed and open multi social-network context allows people to interlink and unify the various URIs that represent themselves. To do so, people can reference a main FOAF URI which can be described via a hand-crafted or automatically generated FOAF profile which links to other existing profiles (and also to interlink distributed social networks from various platforms), as the following snippet and Fig.1 describes:

```
:me owl:sameAs flickr:33669349@N00 ;
    owl:sameAs twitter:terraces ;
    owl:sameAs facebook:foaf-607513040.rdf#me .
```

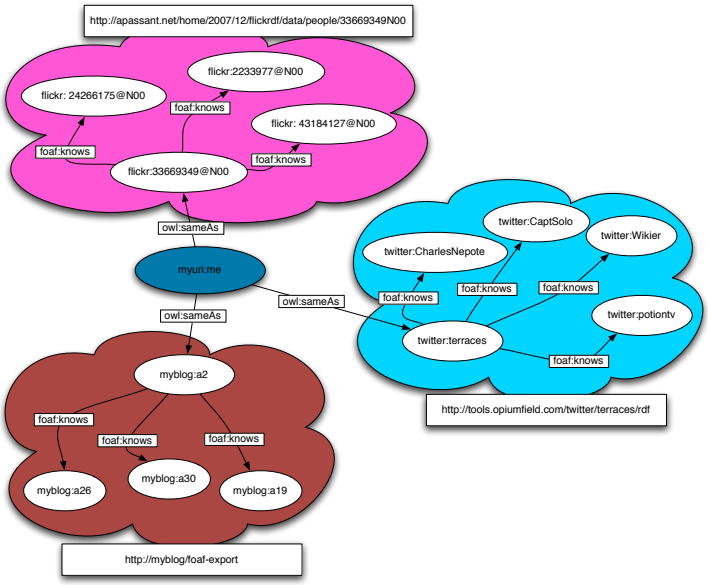

Fig. 1. Interlinking social networks with the Semantic Web

Providing such an entry point allows any RDF-compliant tool to browse one's complete social network in a simple way, i.e. retrieving relationships from Flickr, Twitter or Facebook (1) with standard libraries and SPARQL queries and (2) without having to crawl the Web for data since everything can be accessed from one FOAF file. As an example, we provide a simple script that renders a users' complete social network in a user-friendly Flash interface[11]. This tool

[10] http://www.w3.org/DesignIssues/LinkedData.html
[11] http://apassant.net/home/2008/01/foafgear

only requires the main URI of the user, and thanks to the interlinkage properties described before, it retrieves other URIs and related social networks to render it. This application, which requires only a few lines of Python and SPARQL queries to parse the complete network, shows the benefits of using common semantics to describe networks on social media websites.

Finally, another way to identify uniqueness of a resource across various networks is to implement "reference by description" by identifying some properties that can act as indirect identifiers and asserting that two RDF descriptions using the same value of this property are "talking" about the same entity (using `owl:InverseFunctionalProperty`). FOAF uses several properties of this kind, for example, `foaf:mbox_sha1sum` (i.e. scrambled e-mail) and `foaf:openid` (i.e. an OpenID URL). Thus, even if people use different screen names on various websites, as soon as they register with the same email or with the same OpenID URL, we can uniquely identify then across distributed social networks.

4 Social Network Portability with FOAF

We will describe the use case of a FOAF-aware social media website. Bob, a new user, wants to join a (fictional) Networkr service in order to share some pictures and posts with his friends. To do so, he creates an account using his OpenID URL. While the primary advantage with OpenID is that he does not need a new login and password to connect, it allows the site to easily discover his FOAF profile and URI. Indeed, Bob has delegated his OpenID to his own domain name and has added a FOAF autodiscovery link[12] to his homepage to let software agents easily discover the location of his FOAF profile.

The system will then retrieve Bob's profile as well as his URI (using the `foaf:openid` property) and read Bob's social networks to check if any people in one of his existing networks are already registered on Networkr. The service will then ask Bob if he wants to consider all those people as friends on Networkr. Since Bob does not want to grant access to everyone about his activities on this website, he decides to check himself who to add from his existing friends. All of these steps have been efficiently achieved by the website since it just has to query a single profile and the related RDF description of Bob's social graph. Moreover, each time he logs in, Networkr again browses Bob's complete network to retrieve updates and change local access rights if needed. Thus, as soon as Bob adds someone as a friend on Flickr, he will gain access to his new picture gallery. Finally, since Networkr is completely open and consider that the data and social graph belongs to the user, it allows Bob to export his new restricted network, which he can then reuse on other websites.

[12] http://wiki.foaf-project.org/Autodiscovery

5 Data Portability with SIOC

5.1 Describing Social Media Contributions using SIOC

The SIOC initiative was initially established to describe and link discussion posts taking place on online community forums such as blogs, message boards, and mailing lists. As discussions begin to move beyond simple text-based conversations to include audio and video content, SIOC has evolved to describe not only conventional discussion platforms but also new Web-based communication and content-sharing mechanisms.

As a good Semantic Web citizen, SIOC reuses and extends existing ontologies such as Dublin Core and FOAF in order to be compatible with RDF data modeled using other existing vocabularies[13]. Thus, SIOC is often used in combination with the FOAF vocabulary for describing people and their friends, and the Simple Knowledge Organization System (SKOS) model for organising thesaurus-like data, SIOC lets developers link user-created content items to other related items, to people (via their associated user accounts), and to topics (using specific "tags" or hierarchical categories) - see Fig.2. Through its SIOC Type module, SIOC can represent various types of containers (i.e. `Wiki`, `Blog`, `MessageBoard`) and content items (i.e. `WikiArticle`, `BlogPost`, `BoardPost`).

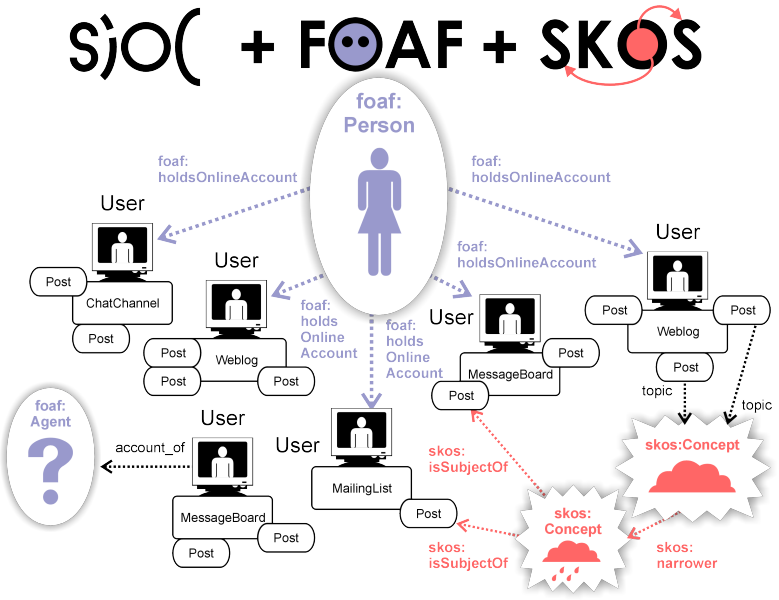

Fig. 2. Combining SIOC, FOAF and SKOS

[13] SIOC Ontology: Related Ontologies and Vocabularies - http://www.w3.org/Submission/sioc-related/

Various tools, exporters and services have been created to expose SIOC data from existing online communities[14]. These include APIs for PHP, Perl, Java and Ruby, data exporters for systems like WordPress, Drupal, phpBB and BlogEngine.NET, data producers for RFC 4155 mailboxes, SIOC converters for Web 2.0 services like Twitter and Jaiku, and usage in commercial products including Talis Engage and OpenLink Virtuoso.

As an example, a SIOC exporter plugin for a blog engine would create a SIOC RDF representation of every blog post and comment, including information about:

- the content of a post (`sioc:content`)
- the author (`sioc:has_creator`)
- the creation / update date (`dct:created` / `dct:updated`)
- tags and categories (`sioc:topic`)
- all comments on the post (`sioc:has_reply`)
- information about the container blog (`sioc:has_container`)

All of these data sources provide accurate structured descriptions of social media contributions (SMCs) that can be aggregated and reused.

5.2 Importing SIOC Data, with a WordPress Example

The SIOC import plugin[15] for WordPress blog engine is an initial demonstrator for social media portability using SIOC. Once installed, it displays a SIOC import panel in the WordPress administrator user interface (Fig.3) which allows a weblog maintainer to import user-created content from another website (described in the form of SIOC data) to their weblog.

Data to be imported can be created from a number of different social media sites using SIOC export tools (described above) which are at the data creation side of the SIOC "food chain" [4]. Since the only requirement for the importer is to "understand" data modeled with SIOC, it makes no difference wether the data comes from another blog, a bulletin board or your latest updates on Twitter. Data sources can also use different ways how SIOC+FOAF data are represented, such as embedding data into web pages with RDFa, extracting data using GRDDL or using external metadata files.

The import process performed by the plugin consists of parsing RDF data (e.g., using open source parsers such as ARC[16] or RAP for PHP), finding all SIOC items to be imported (e.g., all instances of`sioc:Post`) and creating new content on the target site using WordPress API calls. Our "proof of concept" implementation of SIOC importer exploits the interlinked nature of SIOC data and allows to import blog posts along with all the comments that these posts have. To achieve this, the plugin uses `sioc:has_reply` property to identify, retrieve (i.e. fetch additional SIOC data) and re-create comments associated with the blog posts imported.

[14] http://rdfs.org/sioc/applications/
[15] http://wiki.sioc-project.org/w/SIOC_Import_Plugin
[16] http://arc.semsol.org

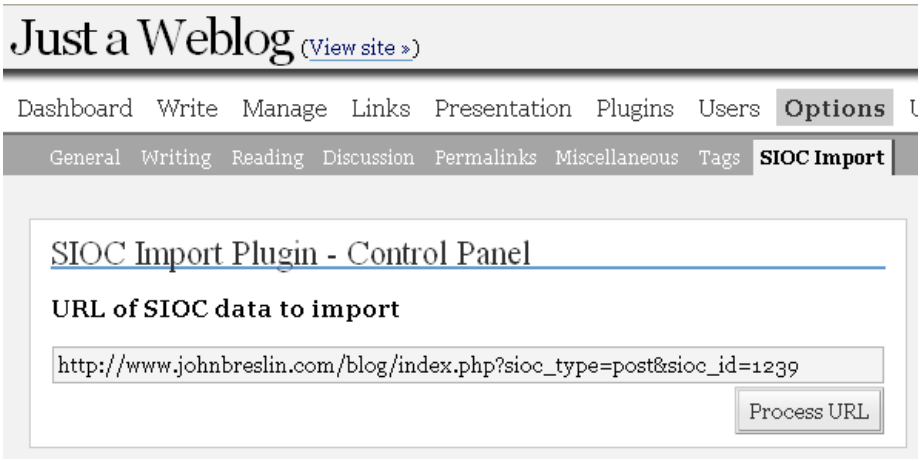

Fig. 3. Importing SIOC data in WordPress

This implementation already allows us to move content between different social media sites. The data format used allows to describe different kinds of information about user-created content such as author information, categories and tags, etc. It can easily be extended to describe additional properties and data types as necessary while still being able to parse SIOC data using a generic RDF parser. As a result, more complex tools for porting data can be built following the same architecture described here.

Horizontal scaling of data import process can be quite simple. In order to port all the data on a social media site (or all the data created by a single user) the import tool will start data retrieval from the top `sioc:Site` profile or a given user profile page. By using the interlinked structure of SIOC data on a site the importer will first retrieve information about all the forums or blogs on a site, and then the information describing user-created content on these blogs or forums. Data import will be similar as for a single blog post, except for the fact that some higher-level objects (forums, user profiles) may need to be created on the target social media site.

Privacy can be one of the challenges when considering data portability. Often, a part of the content that a user wants to port is not for public consumption (e.g. if a user is porting some personal information between her accounts). SIOC can be used in this case since it, as a data format, does not prescribe that all the information should be public. Similar as with other APIs and data formats, the underlying web infrastructure (e.g., HTTP authentication) can be used to ensure privacy and confidentiality.

6 Conclusion

In this paper we introduced a combined use of lightweight Semantic Web vocabularies - SIOC and FOAF - to represent information needed to develop data

portability applications. We demonstrated how FOAF can be used to represent distributed social networks and to manage multiple online accounts held by a person. We showed how SIOC data can be used to represent and port the diverse social media contributions being made by users on various sites. Both approaches can be merged to retrieve and identify all content objects produced by a single user on various social media sites and services, as well as their friends or social network data, thus leveraging the transfer of both social network and content data between social media sites with help of Semantic Web technologies.

7 Acknowledgments

This material is based upon work supported by Science Foundation Ireland under grant number SFI/02/CE1/I131. Authors would like to thank Dan Brickley, one of the creators of the FOAF vocabulary, for his valuable feedback and suggestions for this paper. Authors would also like to thank Tuukka Hastrup for his work on improving the WordPress SIOC Import plugin.

References

1. Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.
2. Chris Bizer, Richard Cyganiak, and Tom Heath. How to Publish Linked Data on the Web. http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/, 20 July 2007.
3. Uldis Bojārs, John Breslin, Aidan Finn, and Stefan Decker. Using the Semantic Web for Linking and Reusing Data Across Web 2.0 Communities. *The Journal of Web Semantics, Special Issue on the Semantic Web and Web 2.0 (Forthcoming)*, 2008.
4. Uldis Bojārs, John G. Breslin, Vassilios Peristeras, Giovanni Tummarello, and Stefan Decker. Interlinking the Social Web with Semantics. *IEEE Intelligent Systems*, 23(3), May/June 2008.
5. J.G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. In *Proceedings of the Second European Semantic Web Conference, ESWC 2005, May 29–June 1, 2005*, Heraklion, Crete, Greece, 2005.
6. Dan Brickley and Libby Miller. FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/.
7. Brad Fitzpatrick and David Recordon. Thoughts on the social graph. http://bradfitz.com/social-graph-problem/, 08 2007.
8. Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. W3C Recommentation, W3C, January 2008.