Freemix: Social Networking Meets Data

David Wood[1 2], David Feeney[3], Eric Miller[4], and Uche Ogbuji[5]

[1] Zepheira LLC, Fredricksburg, VA, USA 22408
david@zepheira.com
[2] The University of Queensland, Brisbane, Australia 4072
[3] Zepheira LLC, Reston, VA, USA 20190
davidf@zepheira.com
[4] Zepheira LLC, Columbus, OH, USA 43065
em@zepheira.com
[5] Zepheira LLC, Superior, CO, USA 80027
uche@zepheira.com

Abstract. This paper introduces the Freemix platform, a framework for building social networking applications that connect people with data. Freemix provides people working with ”desktop” data (such as spreadsheets, XML collections and small databases) or structured web data (RSS, ATOM news feeds, etc.) a means to publish their data in a common translated format suitable for reuse. Once this data is available, Freemix allows users to create customized views of this data reflecting individual or topical preferences and share these views with others.

Freemix uses Semantic Web technologies for several reasons; as a simple, flexible data model for merging, to allow simple descriptive metadata to be added to a data profile to assist the configuration of presentation options for a data set and to as a means for exposing descriptive scaffolds to Web search engines via embedded RDFa attributes.

The Freemix platform is currently being used in several projects. In this article, we will discuss a couple of these briefly to help demonstrate the value of this platform. Future work currently under development will also be discussed.

1 Introduction

This paper introduces the Freemix platform, a framework for building social networking applications that relate people with data. Freemix provides people working with ”desktop” data (such as spreadsheets, XML collections and small databases) or structured web data (RSS, ATOM news feeds, etc.) a means to publish their data in a common translated format suitable for reuse. Once this data is available, Freemix allows users to create customized views of this data reflecting individual or topical preferences and share these views with others.

The Freemix platform is designed to facilitate data dissemination and display in arbitrary social situations. Social connections between people and groups of people may be formed organically, as in existing social networking applications. The platform is aimed primarily at the dissemination and display of data created and stored by single individuals using commodity tools such as spreadsheets and small databases (”desktop” data).

The display and dissemination of desktop data are problems with both social and technical aspects. Information is created in many formats, with many tools, by many people and for absorption by (generally) other people. Often the means of creating and sharing information are technical in nature (such as with a spreadsheet and electronic mail), but the means of absorption remain fundamental: A receiving person has to read the data in order to understand it.

Reading data in the same proprietary program used to create it, as we often do with spreadsheets, belies the obvious fact that this may not be the best way for a receiving person to understand it. Although WYSIWYG word processing programs have addressed display issues for the printed word, little analogous progress has been made for data. Spreadsheets and simple databases have some plotting features, but their use is relatively rare. Only institutional database operators can afford the complicated business intelligence and reporting software necessary to make sense of large-scale relational data. Small-scale database users often rely on simple tabular result formats. Even when reporting options are available, they are not always used.

Freemix attempts to provide an environment where desktop data may be viewed in more than one way by more than one person. Data is converted into a canonical format, augmented with descriptive metadata and made accessible via a World Wide Web URL. Once so organized, multiple views of that data may be readily created by one or more people. End users are able to see their data their way, and also see others' data their way. Mass customization of the display of digital data is thus possible.

Dissemination of desktop information is equally fraught with problems. Spreadsheets and small databases are all too often passed by value, e.g. via electronic mail, instant messaging or file-sharing systems, instead of by reference. Passing information by value naturally results in a proliferation of copies and hence a version control problem. Who has the canonical copy? Do I have a current copy? How can I find out? By contrast, passing data by reference (e.g. by URL) avoids such problems. Freemix allows the passing of data by reference by assigning URLs to not only data, but different views of that data.

Freemix is intended for use by individuals and is for users of all levels of technical expertise who need to share, analyze and access their data. Freemix uses the World Wide Web as a technical platform for both its flexibility and resilience in connecting services as well to gain the widest possible reach.

As wikis and blogs gave everyone the ability to publish their text-based content on the Web to enable sharing and collaboration with others, Freemix allows you to do the same with your data. It allows you to easily mix and merge private and public information, style it with templates and share it with other people in a way they can most easily absorb this information.

Freemix combines two approaches to information sharing developed in recent years. A social network infrastructure allows users to find each other, create linkages and form groups of common interest. Information gleaned from desktop data may be rapidly exposed on the Web, navigated and viewed in various ways. Information publishing is enabled by a builder application that produces views based on the MIT Simile Project's Exhibit application [1]. The combination of rapid information publishing and social network capabilities allow creators of data to more efficiently share that data with others.

2 Related Work

A number of commercial companies and Open Source Software projects have announced facilities for the publication of data to the Web. These include Google Docs, Google Fusion Tables, Oracle Websheets, Anzo for Excel from Cambridge Semantics, and Lyza from Lyzasoft,

Google Docs provides means to publish traditional "office" documents such as word processing documents, spreadsheets and presentations to the Web. Files may be imported or created natively in Google Docs. Google Fusion Tables is a supplementary offering aimed at larger data sets. Fusion Tables is a database and offers database-like functionality such as filtering and merging of data. Visualization of data in charts, graphs and maps is possible.

Oracle Websheets allows users to upload spreadsheets onto the Web, edit them on the Web, create reports and share them with others via a permissions-based authorization scheme. Spreadsheets can also be exported from Websheets back to their original spreadsheet format.

Anzo for Excel is a commercial product to publish spreadsheets and data from databases onto the Web, and merge their contents. Users semantically type data elements to facilitate multi-author sharing with meaning.

Lyza is a business intelligence project aimed at corporations. The product allows data from various sources to be combined for analysis. The interface to Lyza is spreadsheet-like. Users may browse joined set of data and create custom reports.

The Open Source content management system Drupal supports the creation of Simile Exhibits and Timelines for the display of data.

Although some of these approaches, particularly Google Fusion Tables, Oracle Websheets and Anzo, facilitate data sharing between people who already know each other, none of them were designed for the bottom-up identification and reuse common in social network approaches. Freemix is an attempt to present data for reuse within the context of a social network in which relationships may be established organically within the technical environment.

3 Extending Social Networks for Data Concerns

Data sharing and quality has traditionally suffered for social reasons. Data is locked away on user desktops or in enterprise databases and hidden from those who need to use it or who have information which may help improve it. It is typically provided to users in views that the data provider believes is useful which may bear no relation to actual user needs. There is little opportunity to have a dialog about the data.

Freemix extends the concept of social networks to data in order to allow the community to unlock its potential. The community is empowered to create new views which allow them to explore and evolve that data based on their collective knowledge. The ”tribes” feature allows for the creation of communities around common areas of interest. Data, views and discourse relevant to these common interests can be shared within the tribe. Tribes promote discovery of new information based on new views, to augment data by connecting it with other data sets, and to improve the quality of the data via the shared knowledge of the community.

Once limited to views of data in columns and rows or as pushed to them in static reports, users of Freemix can view a data set in a variety of new ways based on what makes sense to them. Once a user indicates what types of data are present in their data set, Freemix allows them to easily plot on a map, on a timeline, or in a variety of other ways as makes sense from the contents of the data set. Viewing the data in these new ways leads to new and valuable insights which can be easily shared with others.

Freemix tribe members can discover data sets provided by others which, when connected to their own data set, may enhance its value. For example, imagine a tribe whose area of interest is gardening. One user may have a personal spreadsheet of vegetables which they've been thinking about planting that year. Another community member may have access to a complete listing of vegetables along with their suggested planting dates and gestation periods to grow. Merging these two data sets would provide new insights into when and how best to structure and plant ones garden. The fruits and vegetables from ones garden can then in turn be merged with another Tribes favorite recipes to help narrow in on the best meals that match the ingredients ready for harvest. There are an open ended set of possible options for stitching together data. The Freemix platform is designed to reduce the cost in allowing such exploration and sharing.

Data curation has typically been a task assigned to a handful of expert users who have been given special permission to add to and modify it. Centralized control of data has been seen as essential to maintain data quality, but is often an impediment. By opening up the data, Freemix gives the community a vested interest in improving the data as well as a forum in which to communicate these improvements. Providers of the data are incented to ensure a higher level of quality as they know their data will be viewed and shared with many. Users of the data are incented to communicate errors they find back to the data owner since they want the data they rely on to be accurate.

4 First Class Objects and Relationships

Freemix attempts to equally and symmetrically expose four types of objects to the world; users, groups of users (”tribes”), data profiles and views of data. Like any social network, users and groups of users figure prominently. The purpose of Freemix is to facilitate the sharing and viewing of data, so data profiles and views of data are also prominent. We call the four objects ”first class” because the relationships between them define the non-administrative workflows of the environment. Figure 1 summarizes the relationships between Freemix's first class objects.

A data profile is a logical concatenation of a data representation and metadata describing it. A data representation is a representation of a user's original data in a canonical format (we are currently using Javascript Object Notation, JSON, as an implementation convenience). Original data is converted to JSON via an external data conversion service based upon Akara [2]. A descriptive title for the data file is collected from the user upon upload. Additional metadata is collected and used for the purpose of enhancing display options for views built upon the data, as described in the section Semantic Augmentation of Data Profiles, below.

Data profiles may have multiple views built upon them. That is, if a user A uploads a data profile, both user A and other users may use the same data profile to create (probably different) views of the data. Each user has the opportunity to view the data as they see fit, in the way that makes the most sense to them. For example, user A may create a view that plots location-based data on a map and time-based data on a timeline. Another user, call them user B, may wish to pull numerical data from the same file and plot it on a scatterplot or another view appropriate to numbers. Multiple views can exist side-by-side on a given Web page.

First class objects have natural relations to each other. A view, like a data profile, is created by a user, who in turn has relationships with other users (as friends, a direct connection, and/or tribes, allowing connections relating to a particular interest). Views and data profiles may be published to a tribe so tribe members can share the information. Similarly, looking at a view or a data profile informs a user which tribes have the information published to them. Relationships between first class objects are intentionally bidirectional and symmetric.

5 Human Annotation and Semantic Augmentation of Data

An important design criterion for Freemix was to facilitate multiple and differentiated views of data profiles. Data may be displayed in lists or tables, plotted on maps, timelines, scatterplots, pie charts, etc. Semantic metadata is collected at design time for views and used to facilitate available view stylings. If location data is present, say, then a map may be appropriate. If temporal information is present, then a timeline view may be appropriate. Numerical data may suggest plots of various types. As of this writing, semantic metadata in data profiles are used only to facilitate the configuration of a particular view once a user manually chooses it.

Unfortunately, the lack of standardized input formats for common data types, such as dates, times, and locations makes it difficult to guess a user's intent. Al-

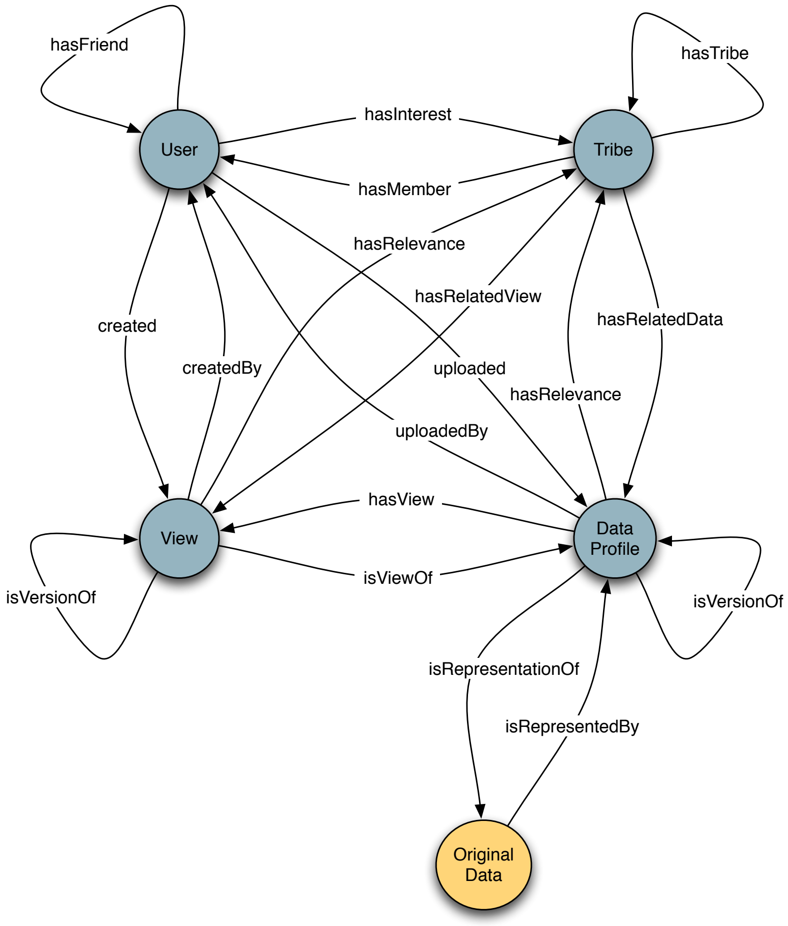

Fig. 1. Freemix First Class Objects

though it is certainly possible to address this type of problem technologically, an underlying premise of Freemix is that users know more about data than machines. Freemix instead approaches the assignment of semantic metadata socially: A user is presented with an interface upon data upload that allows the user to manually assign data types to data elements. Humans are simply better at this task than computers using current technologies.

For example, consider some input document that contains dates. The date ”December 11, 2009” may be represented as ”12/11/2009”, ”11/12/2009”, ”11/12/09”, ”11Dec2009” or any of many other possible variations. Regional differences in month-day order are particularly vexing for machines, but are recognized quite efficiently by humans. Now consider input documents that refer to geographical places. The strings ”Columbus, OH” or ”6565 Franz Road” are arbitrary examples of references to place. These again are quite efficiently known as geographical locations by the humans as context (in the case of reading this prose) has been established. This context is critical and readily understood by curators of the data (or even in some cases casual users of data) but again difficult to determine by a machine. Allowing users to help assert this context and make it explicit enable the Freemix platform to act on it. An assertion of ”date” for example allows the user to choose services to normalize dates accordingly. An assertion of a ”place” allows for geospatial coordinate lookup to occur. The ’types’ of information a user can assert is customizable along with the availability of corresponding services. Freemix allows, however, these assertions to be re-used by others thus building off the collective understanding of others.

Table 2 summarizes the default data types that may be associated with data elements. All data elements are presumed to be of type ”text” by default. Selection of the ”url” type results in a data element being wrapped in a hyperlink tag pointing to the data element, which is presumed to be a valid URL. URLs that resolve to images may be typed as ”image”, in which case the image, not the URL, will be displayed inline wherever it is to be shown. Dates, datetimes and locations are used as clues in the configuration of map and timeline views.

RDF metadata describing data profiles is currently stored in a JSON serialization format, not in RDF/XML. The Freemix team is still experimenting with the most appropriate way to store and use this metadata to greatest effect.

Table 1. Semantic Descriptions of Data Elements

Type	Results in	Used with
text	Inline display without modification	All Views
url	Hyperlinks	All Views
image	Image content	Table, List Views
datetime	ISO 8601 dates or datetimes.	Timeline View
location	Latitude and Longitude	Map View

6 Exposure to Semantic Search

Generated views of data are bootstrapped from a small amount of HTML dynamically served via a template. The HTML refers to Javascript libraries, CSS stylesheets and the underlying data. Unfortunately, the minimal HTML scaffold is simple enough that it does not provide adequate information for search engine crawlers to index the generated content in a meaningful way. Difficulty of indexing simple HTML scaffolding is a recognized problem for others using Exhibit [3].

The Freemix platform has addressed search engine indexing by means of RDFa [4]. RDFa markup is added to generated HTML scaffolding at resolution time for a view's URL. Support for RDFa indexing by major Internet search engines is a rather new phenomenon. The two largest search engines, Google and Yahoo!, have recently announced support for both indexing content including RDFa and building search results based on RDFa information [5], [6].

Table 1 summarizes the descriptive metadata exposed about a view using RDFa. A view's author (or creator) is related to the view using a Dublin Core Metadata Initiative ”creator” attribute. The author may be further described using information from that user's profile information, including given and family names, affiliations, location and even an identifying photograph or icon. The Friend of a Friend (FOAF) and Google's People vocabularies are used for detailed descriptions of authors.

A view's title and descriptive abstract are also provided via Dublin Core relationships. Abstract information is extracted from textual descriptions of a view if such a description has been provided by the author. Publisher identification is set on a site basis and related via a Dublin Core attribute.

Both views and their underlying data may be licensed. In our early implementation, variations of Creative Commons licensing are allowed.

A view is guaranteed to have an author, title and publisher. Detailed author information and abstract information are not guaranteed to exist for any given view. Discussions regarding the enforcement of content licensing are ongoing.

Table 2. Supported RDFa

Concept	Vocabulary
Creator	Dublin Core, FOAF, Google's People
Title	Dublin Core
Abstract	Dublin Core
Publisher	Dublin Core, Google's Businesses and organizations
License	Often Creative Commons

In addition to the metadata about a view, the Freemix platform also provides an effective means for allowing the raw data associated with any view available as RDFa. The data captured from the users ability to annotate the kind of data that is available is also added to the RDFa stream providing richer means for searching and discovering the raw data addressing some of the indexing issues identified using Exhibit [3]. Further, this capability, when coupled with appropriate CSS techniques provide the basis for making the underlying data to Exhibit more compliant with Web Accessibility Initiative (WAI) guidelines.

7 Early Implementation Experience

There are an increasing number of communities using the Freemix platform to more effectively share and discover data. Many of these examples are found inside organizations but increasingly more are public.

One of the initial public efforts that is leveraging the Freemix platform is The National Digital Information Infrastructure and Preservation Program (NDIIPP) of the US Library of Congress.

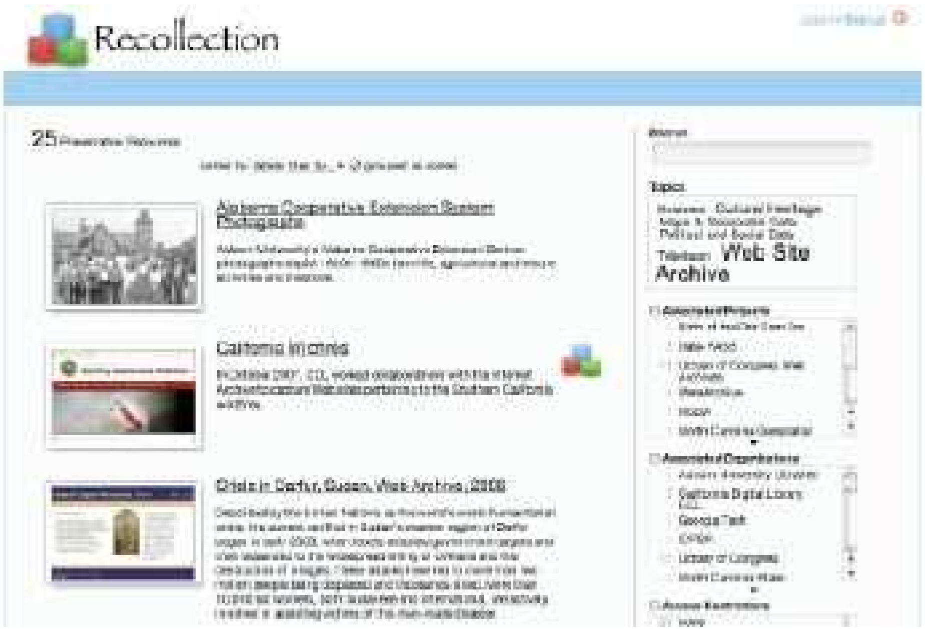

Fig. 2. Recollection Pilot Using Freemix

The NDIIPP program is an effort to develop a US national strategy to collect, archive and preserve ”at risk” collections of digital content for current and future generations. It is based on an understanding that digital stewardship on a national scale depends on active cooperation between communities in public and private sectors. The Library has built a preservation network of over 180 partners, both in the US and internationally, to tackle the challenge, and is working with them on a wide spectrum of initiatives including collections of historical, scientific, cartographical, media, legislative and sociological materials.

In 2008, the NDIIPP partners shared content through a simple web page. In order to explore more useful tools and processes for sharing diverse content across partners collections, the Library began a pilot project in 2009 with Zepheira, LLC to develop a proof of concept that can be used to collect and explore information about digital collections. The result is a software platform called Recollection. Freemix was in turn chosen as the basis for Recollection.

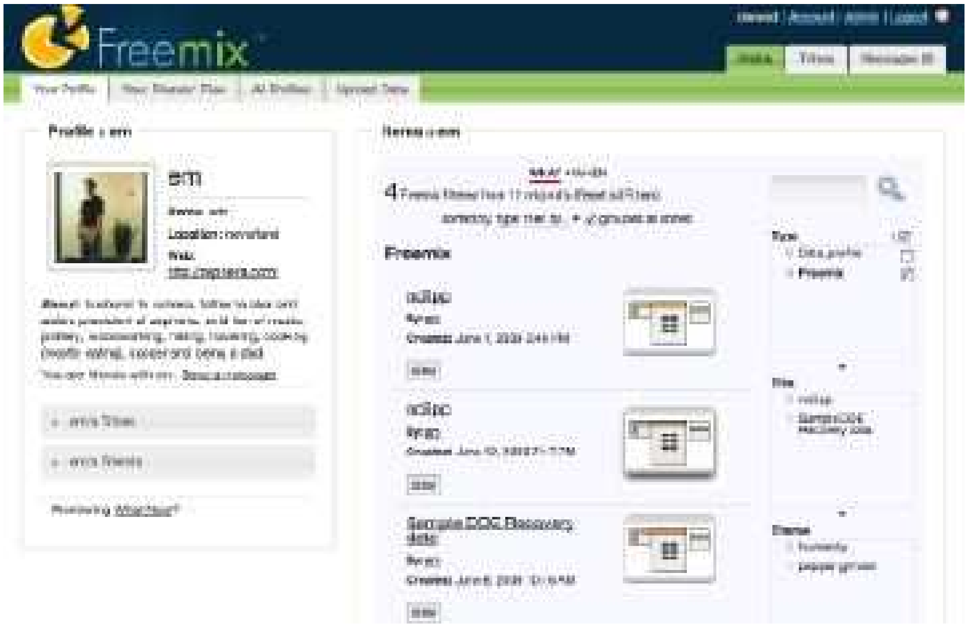

Fig. 3. Freemix.it Service Using Freemix

The focus of the Recollection Pilot has been to explore how to use the Recollection tools:

1. to increase the ability to access and connect information in diverse digital collections by exposing the metadata and content in different ways; and 2. to investigate methods for processing the metadata and content to make them available for the platform.

Recollection allows communities of interest to interconnect their information. The pilot platform uses semantic technologies to enhance discoverable access for NDIIPP collections, making them easier to find, access, and share, and integrate with other digital information sources. The platform enables third-party applications developed by private or public organizations as well as interested individuals to support education, research, policy analysis and other completely unforeseen uses. The Library was particularly interested in the framework's ability to facilitate the identification, location and reuse of information in NDIIPP collections, and the ability to provide an open interface for third parties, to plug services and applications into that framework. The Library and Zepheira plan to continue their collaboration to build out the Recollection vision on the Freemix platform.

The Freemix platform has also been used to field a free Internet-based public service aimed at the general public. Freemix.it is a public social network for individuals with data to share. Freemix.it is in an invitational Beta testing period as of this writing and is anticipated to launch for the general public in early 2010.

8 Future Work

Not all symmetric relationships between first class objects are currently implemented. One, the relationship noting that an original datum is represented by a particular data profile, is unlikely ever to be implemented due to technical constraints existing in proprietary file formats and a lack of control outside the scope of Freemix. Versioning of views and data profiles is incomplete, but is expected to be implemented fully in due course.

Freemix's use of Akara to provide data transformation services allows for rapid support of new data formats. However, central management of available data transformations is currently enforced. It would be useful for end users to have the ability to extend data transformation capabilities for their own use and for others to be able to reuse them. Some form of data transformation marketplace may be needed.

Privacy and access restrictions have not been addressed. Early users of Freemix have either been publishing public information (NDIIPP) or forced to limit themselves to information they wish to make public (freemix.it). Business and government users have expressed interest in access control mechanisms.

The ability for users to easily mix multiple data sources (at both property and value space) is currently an active area of work and expected to be made available later this year. Mixing of arbitrary data is technically challenging. The Freemix team is attempting to avoid some difficult issues such as identity resolution, similarity joins and schema mapping via the presence of a human in the loop. Similar to the approach taken by Cambridge Semantics' Anzo for Excel, users are responsible for semantically tagging data during a merge in order to both type the data and to remove name and schema conflicts.

Data profile information such as data types are currently used to facilitate the configuration of different views, such as a map or a timeline. Future use of that information could include the offering of appropriate view types to a user. If the data includes location information, for example, a map view may be rendered by default.

9 Conclusions

Freemix provides a useful intersection of social networking practices and data manipulation. This paper provided a brief overview of the Freemix framework, discussed some of the Semantic Web approaches used by Freemix and noted two Web-based projects currently using Freemix.

Indexing of small HTML scaffolds would have been difficult or impossible without using a metadata-based approach. We were fortunate that the Internet search engines with the largest market share decided to support RDFa at the time we needed it. Most data uploaded by Freemix users does not contain adequate descriptive text to allow search engine indexing by traditional means.

Gathering of data typing information allows Freemix to have a smaller user interface and allows the user to create complicated views on data with fewer interactions.

10 Acknowledgements

Our thanks to the following Open Standards, Open Content and Open Source Software communities for providing us critical building blocks, ideas and support: The World Wide Web Consortium for RDF and RDFa, Creative Commons for copyleft, The MIT Simile Project for Exhibit and Simile Widgets, OCLC and the PURL Project for PURLs, The Akara Project for the Akara data transformation pipeline, the Django Project [7] and the Pinax Project [8] for the Django Web framework and related applications.

References

1. MIT: The mit simile project. http://simile.mit.edu/ (2009)
2. Ogbuji, U.: Akara. http://xml3k.org/Akara (2008)
3. Huynh, D.: Google's rich snippets in exhibit. http://tinyurl.com/huynh-googlerichsippets (May 22 2009)
4. Adida, B., Birbeck, M.: Rdfa in xhtml: Syntax and processing. http://www.w3.org/TR/xhtml-rdfa-primer/ (October 2008)
5. Goer, E.: Searchmonkey support for rdfa enabled. http://tinyurl.com/yahoosearchmonkey-rdfa (September 2008)
6. Google: Marking up structured data. http://tinyurl.com/qthorl (May 12 2009)
7. Django Foundation: Django — the web framework for perfectionists with deadlines. http://djangoproject.com (2009)
8. Tauber, J.: Pinax — a platform for rapidly developing websites. http://pinaxproject.com