Talking about Data: Sharing Richly Structured Information through Blogs and Wikis

Edward Benson, Adam Marcus, Fabian Howahl, and David Karger

MIT CSAIL
{eob,marcua,fabian,karger}@csail.mit.edu

Abstract. Several projects have brought rich data semantics to collaborative wikis, but blogging platforms remain primarily limited to text. As blogs comprise a significant portion of the web's content, engagement of the blogging community is crucial to the development of the semantic web. We provide a study of blog content to show a latent need for better data publishing and visualization support in blogging software. We then present DataPress, an extension to the WordPress blogging platform that enables users to publish, share, aggregate, and visualize structured information using the same workflow that they already apply to text-based content. In particular, we aim to preserve those attributes that make blogs such a successful publication medium: one-click access to the information, one-click publishing of it, natural authoring interfaces, and easy copy and paste of information (and visualizations) from other sources. We reflect on how our designs make progress toward these goals with a study of how users who installed DataPress made use of various features.

1 Introduction

Recent efforts to generate and curate high-value structured datasets have made great headway on several fronts, as exemplified by open government initiatives, Facebook's Open Graph project, and Freebase's structured wiki. While these centralized, top-down approaches are significant, we have yet to see wide adoption of structured data publication at the grass-roots level. Taking note that the development of hosted blogging platforms encouraged millions of web readers to become content authors as well, we aim to entice these users to publish data by building data-oriented features into their existing blogging software.

Large projects can rely on the promise of societal and technical benefits to justify the costs required to curate and publish structured data. We believe that for independent bloggers to take part in data publishing efforts of their own, the promise of later portability and reuse is not enough. Instead, end-user-focused data publishing tools should offer immediate gratification in the form of useful visualizations and interesting data aggregation before they focus on formal ontologies and namespaces. Only after the user has seen the benefit of data publishing as part of their content authoring workflow can we take steps to link, integrate, and further reuse the underlying data.

We also take inspiration from efforts such as the Semantic MediaWiki project, which has brought structured data publishing to wikis by exposing it in the WikiText format already familiar to wiki users. We aim to similarly provide bloggers with data publishing tools that blend in with existing blogging environments. The popular blog publishing platforms that we target differ from wikis in that they depend more heavily on WYSIWYG editing, click-to-embed rich media, and an easy-access copy-and-paste culture. To facilitate the adoption of grass-roots data publishing, we must build tools that minimize the difference between traditional text-based blogging and the future of publishing, in which all content producers are data publishers.

To understand how to accommodate the data-blogger of the future, this paper first examines the properties of blogging platforms that led to their popularity among content authors on the web. We then demonstrate a latent need of, and great potential for, data-centric blogging tools with a content study of 210 blog entries on the web. This study quantifies the kinds of data-supported arguments that blog authors make and shows that bloggers are already using structured data in their content, but the tools they have to communicate it are limiting.

We then present DataPress, a plugin for the WordPress blogging platform which facilitates data visualization from minimally structured files, allows bloggers to point at other data presentations as a starting point for their own, and allows bloggers to publish and aggregate their own data sets. We build into DataPress the ability to easily link to external data from spreadsheets, RDF sources, and Semantic MediaWiki sites. We further demonstrate the ability to rely on Semantic MediaWiki as a community ontology server to encourage schema convergence across data feeds produced by DataPress bloggers.

Finally, we examine the log data and interview the authors of real DataPress deployments, one of which having seen over 55,000 page views. These users provide insight into how web authors are using DataPress to publish data today (not in their blogs, to our surprise) and where opportunities exist to improve data publishing tools for tomorrow.

2 Requirements for the 21st Century Blogger

Our goal is to bring structured data publishing to the blogging community, and to do so we must build tools appropriate to the environment that bloggers expect. This section examines the properties of blogging platforms that made them such a successful grass-roots medium of contribution to the web. We expect that by preserving these traits in a data-oriented blogging tool, we are more likely to gain traction from the blogging community. Of the many features of blogging systems that make them popular, we highlight:

One-click Publishing Though publishing through a blogging platform is equivalent, from a technical standpoint, to uploading HTML documents over FTP, the increased usability and convenience that a web publishing interface provides encourages participation by a far wider audience.

Visual Authoring Environment Blog platforms offer users familiar, word-processor-like WYSIWYG text editors, with HTML forms to guide them through more complicated tasks. Notably, the author does not have to coordinate her work through several distinct applications—all her authoring needs are met within the editing environment of the blogging tool.

Copy and Paste Web blogs have developed a publishing culture that makes significant use of copy and paste, both to quote information found in other sources and also to replicate layout or visualization functionality that the author could not construct herself. Sometimes, such quoting is an end in itself; at other times, the goal is to use the original content as a starting point for publishing by modification. It is much easier to copy someone else's nicely formatted page and replace its content with your own than to understand how to create such a layout in the first place.

Pre-Packaged Widgets Blogging systems make it easy to include rich media widgets—such as slide shows and video clips—in article text without having to manually write code or configuration. By simply uploading several pictures (in the case of a slide show) or adding a link (in the case of a YouTube video clip), the blogger benefits from the platform's ability to package up this simple data into a rich format that entertains visitors.

These traits help blogging platforms turn the technical task of publishing web content into an easy process accessible to the grassroots authors that provide so much of the web's "long tail" of content. If we wish to encourage these grassroots authors to provide data as well, we must give them tools that work from within these familiar environments and share their properties.

We use these traits as a guide to construct DataPress, and we argue that as a result DataPress can support, for rich structured data, the same behaviors that made the web's text authoring tools so effective:

- It has the same "click and you see it" immediacy that made hosted blogging such a big change over the FTP publishing workflow: it enables users to insert data into a blog post the same way they insert an image, offers readers embedded data visualizations inside article bodies, and it does so without leaving the metaphor provided by the blogging platform
- It does not require the author to understand complex data models, but instead can be based on concepts already familiar to end users: simple forms, embedded media, and links to data-laden websites
- It offers the same copy-and-paste workflow as text, making it easy for authors to "quote" both the data and the data visualizations authored by other users, either to be used unchanging or as a starting point for authors (who may not yet know how to author their own data or visualization) to make their own points by authoring changes in the acquired data or visualizations
- It inspects a user's data in order to better guide her through the creation of rich, interactive visualizations. Users can add faceted navigation, interactive maps and timelines, and search functionality all by selecting a few options in the blogging editor.

3 The Latent Potential for Grassroots Data

Blogging platforms facilitated the enormous growth of the web over the past two decades, but the capabilities of these tools are primarily limited to text. In contrast, professional publications often deal with rich, structured data: shopping sites offer faceted browsing across their product databases; product review sites let readers dynamically pit products against each other in feature-by-feature tables; and news sites such as the New York Times (which runs its own Visualization Lab [1]) publish interactive presentations of complex information. Arguably, these professionally managed web sites are significantly more expressive than grassroots authors' pages. One might think that this is because only large professional publishers care for such expressivity, but we observe that the desire to publish and present data extends far beyond large publishers.

In this section we present the results of a blog content study that indicates that bloggers are in fact *already* frequently talking about data; they are just doing it using text and static images, the best way that they can given their current publishing platforms. We believe this is a hopeful result for the semantic web community, for it suggests that grassroots bloggers would be eager to make use of structured data if their tools made this process easy and beneficial to their needs.

For the purposes of this study, we use the term *blog* to refer to any article-style publication on the web, including both personal journals and professional periodicals. A *semantic entity* refers to an object with one or more properties described in structured or unstructured (natural language) form. A *collection of semantic entities* refers to a sequence of semantic entities of the same type described in a document. For example, a semantic entity might be a paragraph of text or a table row that describes the technical specifications of a new camera. A collection of semantic entities would be a text document or full table comparing several cameras to each other.

We coded 210 blog articles across 21 blogs to measure the occurrence and nature of semantic entities and semantic entity collections within their text [2]. We generated this blog sample by selecting the 10 most recent entries (at time of study) from a semi-random list of blogs taken from the Technorati[5] blog indexing service. This list of blogs included:

- The top ten blogs according to Technorati's "authority rank"
- Eleven blogs selected at random from Technorati's list of "rising" articles

We used this selection method to attempt to capture both high quality, professional content (top ten blogs) and also blogs that varied in style and represented the "long tail" of the web (top rising posts). For each blog in our sample, we downloaded its RSS feed and coded each of its ten latest entries.

Aggregating across the articles for each blog, we found that:

[1] `http://vizlab.nytimes.com/`

[2] The data for this survey can be found at
`http://projects.csail.mit.edu/datapress/content_survey`

	Lone Semantic Entities	Collections of Semantic Entities	Visual Collections	Referenced Datasets	Referenced Resources
Articles with one or more occurrences (of 210)	45 (21%)	64 (30%)	22 (10%)	67 (32%)	191 (91%)
Total Count	58	105	49	428	1061

Table 1. Number of occurrences overall and number of articles with various properties.

- 17 of 21 blogs contained at least one article in their latest 10 that enumerated the properties of a single semantic entity.
- 18 of 21 blogs contained at least one article in their latest 10 that enumerated the properties of a collection of semantic entities.
  - Half of these blogs used natural language text to describe the collection.
  - The other half used a table or a static image containing an info-graphic.

Aggregating across the articles for *all* blogs, we found that:
Table 1 shows us that:

- 21% of articles surveyed contained at least one semantic entity.
- 30% of articles surveyed contained a collection of semantic entities (anecdotally, these were things such as polling results in different states, economic conditions in different countries, and professional sports records).
  - Two-thirds of these collections were presented in natural language text instead of a structured or visual format.

Finally, our data revealed that blog entries frequently refer to external sources of data rather than present original content. Authors made reference to some externally attributed datum or statistic in 91% of articles surveyed. In, 32% of articles, this reference was to an explicit data set, often given by name (e.g., "A 2008 Zogby Poll reported that...''), while in the other 59% it simply referred to a person or organization who had claimed the truth of the numerical fact. In all, we counted 428 total references across the 67 articles which mentioned data sets. These numbers are surprisingly high, perhaps influenced by the fact that our study was done in the midst of an electoral season, but they serve to reinforce the intuition that bloggers are in many respects serving as topic- or geo-localized journalists. They are writing about issues, and these issues involve data. We aim to make that data navigable, linked, and reusable.

Anecdotally, much of the presentation of semantic entities was inlined in text, rather than in a structured tabular format. Interactive data visualizations were rare—structured presentation tended to be either static tables or images. In fact, most of these collections were included in an HTML table or rudimentary list rather than a full-blown visualization. Data "links," if at all present, tended to be narrative references to a data set rather than resolvable URLs.

These results suggest significant latent potential for tools that allow bloggers to publish data with the same ease with which they already publish text. These authors are already interested in data, and at times they are publishing tables or images, indicating that they are willing to adopt non-prose presentation styles if they are available.

Currently the only non-prose data presentation tools that blogging platforms support are HTML tables and static images. We aim to fill this gap with DataPress, which provides both interactive visualization capabilities as well as raw data publishing and linking.

4 DataPress

DataPress[3] is our attempt to create a blogger's tool to publish, share, and copy data and data visualizations. We start from the premise that, from the standpoint of content authors, visualizations are an end in themselves—if a picture is worth a thousand words, surely a good interactive visualization is worth at least tens of pictures. They allow anyone encountering that data to understand it better by exploring it.

But DataPress is also a means to an end: making data more easily available for reuse. DataPress' rich data visualizations encourage authors to use it, but the tool also *exposes* the data it is showing off, making it easy to link to or snapshot, thus enabling the same reuse ecology already pervasive in textual blogs. With this in mind, we will describe DataPress in its four distinct roles:

1. Authoring data
2. Consuming data originating elsewhere
3. Authoring visualizations
4. Exposing data for consumption by other tools

For the authoring roles, our key goal is to fit data and visualization authoring naturally into the already existent workflow associated with WordPress. For the data sharing roles, we arrange for our tool to offer, with no extra user labor, JSON and RDF data "feeds" that can be consumed by others. We also facilitate easy linking of diverse content on the web for aggregation and visualization within a blog post.

DataPress is implemented as a plugin for the WordPress blogging platform. We chose WordPress because it has a large install base (over 3,816,965 downloads in 2007 alone [3]) and because it is a blogging tool used widely for both personal blogs and professional publications, including media outlets as large as the Washington Post online edition. Like other blogging platforms, WordPress places a high value on guided workflow and simple form-based configuration, so the DataPress plugin exposes all of its features as enhancements to the existing WordPress authoring interface.

This section describes the most recent version of DataPress. Our user study, presented in Section 6, was conducted across users of a previous release of DataPress. This previous release contained of all the features described below except for individual item publishing (in Section 4.1), data feeds (in Section 4.4), and Semantic MediaWiki hooks (Section 5).

[3] Downloadable source code, examples, and demo blog for testing available at: `http://projects.csail.mit.edu/datapress/`

4.1 Authoring and Uploading Data

DataPress allows authors to create individual data items to publish with a post or upload entire datasets at a time. Both these option are accessible from buttons added to the WordPress post editor, seen in Figure 1.

Authoring a Data Item By pressing the Data Item button, bloggers can enter key-valued information to associate with a typed semantic entity and publish this information as metadata with a blog post. This usage scenario fits the type of blogger who publishes similarly themed articles over time and would like to benefit from being able to aggregate their structured content for presentation purposes or export to the community.

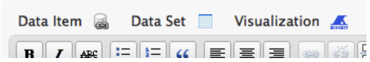
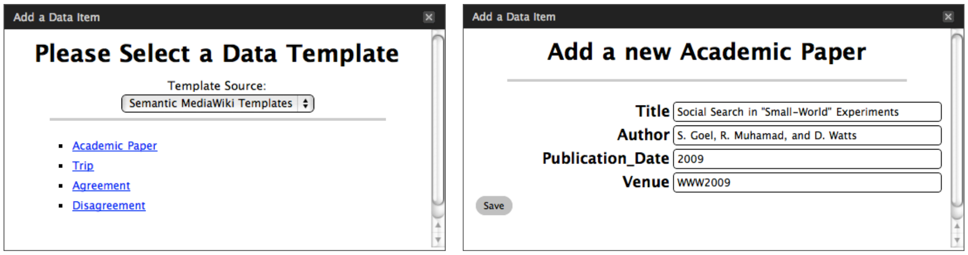


Fig. 1. DataPress Entry Points

Consider the practice of blogging one's academic reading list—some students and professors enjoy blogging summaries of papers they have read so that they can share their thoughts with others in the community. DataPress allows this temporal stream of activity to be published as structured data as well. While writing the blog post, the author clicks the "Data Item" button seen in Figure 1 and DataPress will bring up a selection of "data templates," shown in the first screenshot in Figure 2. A data template is simply a blank form derived from the schema of some item type.



Please Select a Data Template

Template Source: Semantic MediaWiki Templates

- Academic Paper
- Trip
- Agreement
- Disagreement

Add a new Academic Paper

Title	Social Search in "Small-World" Experiments
Author	S. Goel, R. Muhamad, and D. Watts
Publication_Date	2009
Venue	WWW2009

Fig. 2. Choosing and filling in a Template

This list of data templates can draw from a variety of places. DataPress comes with a collection of built-in data templates, such as academic papers, books, and workouts, but it can also be configured to talk to Semantic Media Wiki installations or other data template repositories on the web, allowing the blog author to take advantage of communities that maintain such information and encouraging schema convergence across web sites (this idea will be expanded in Section 5). If no template fits the item, DataPress allows users to create their own by entering a custom class name and the properties that should be associated with its instances.

Once the user has selected a data template, DataPress loads the template schema (possibly from a remote repository) and transforms it into a web form for the user to fill out. One such form is shown in the right-hand side of Figure 2. From here, the data is stored in DataPress' back-end database, while a visual marker for the data is embedded into the text of the post as a small icon that allows editing or removal of the item from the post, shown in Figure 3.

The Data Item interface allows a blogger to follow their natural habit of writing a new article about each data item, while also producing an aggregate data set over time and across blog posts for rich visualization. Our reading-list blogger can place, sticky on their front page, a single rich "My Reading List" visualization showing all articles they have read, with links to the individual blog postings about the articles. This visualization becomes a new, non-chronological index into their blog content.

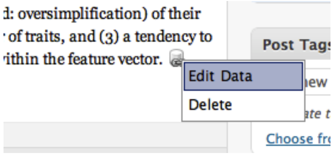

d: oversimplification) of their
of traits, and (3) a tendency to
ithin the feature vector.

Fig. 3. Data Item in a Blog Post

Uploading Data Sets DataPress also lets users associate entire data sets with a blog post. Using WordPress' built-in file upload tool, they can upload a file, and then using the Data Set button provided by DataPress, they can associate that file with a blog post. DataPress utilizes the data import mechanisms of the Exhibit framework [12] and the MIT Babel [8] data translation web service to accept a wide variety of formats, including RDF, JSON, CSV, XML, Microsoft Excel, and Bibtex.

Once a data file is associated with a blog post, DataPress stores this information in its database and provides the option of attaching *Data Footnote* links at the end of the blog post, allowing the reader to visually see links to the data that accompanies the writeup. These associated data sets are also used as inputs for data visualizations, shown later.

4.2 Data Linking

DataPress also lets authors link to remote data sets via URL. In addition to supporting a wide number of data formats that can be linked to directly, DataPress contains a special importer that handles what we call *approximate links*—URLs that point to web pages that talk *about* data, rather than links to the raw data itself. We currently support four such kinds of approximate links:

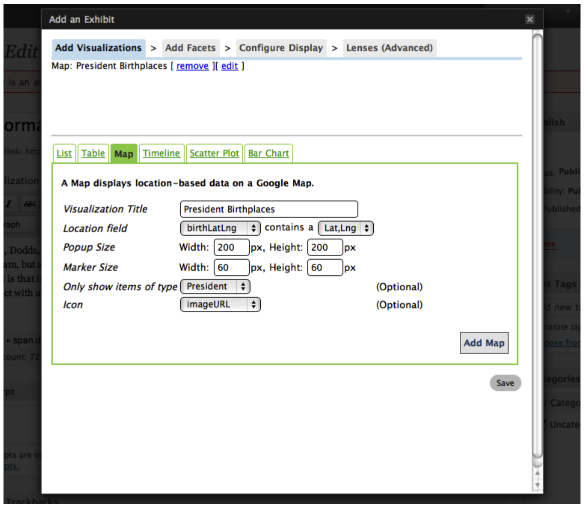
- URLs of DataPress-powered pages are automatically converted into data links to that page's data sources
- URLs of web pages containing an Exhibit-powered visualization are automatically converted into data links to that page's data sources
- URLs of Google Spreadsheet files are automatically converted into API calls into Google's JSON data service

– URLs of third-party JSON data files are converted into JSONP calls routed through a DataPress JSON-to-JSONP service

We expect to grow support for approximate linking as we believe that it supports, for data, the same copy-and-paste-ability that made blogging tools successful. If a user sees a page with data they want to use, they should only have to copy and paste that page's URL to be able to remix and republish its data. As we will show in the following section, we are currently also working on support for easy import of Semantic MediaWiki data via remote `ASK` queries.

4.3 Visualization Authoring

The "Visualization" button, shown above the post editor in Figure 1, provides access to a wizard which walks the user through the creation of a data visualization. DataPress uses the Exhibit framework for displaying interactive visualizations. This allows the plugin to benefit from the developer community that builds data importers and visualization plugins for Exhibit. DataPress' configuration wizard, shown in Figure 4, contains many of the various options Exhibit provides, as well as some blog-specific enhancements.


Add an Exhibit
Add Visualizations > Add Facets > Configure Display > Lenses (Advanced)
List
Table
Map
Timeline
Scatter Plot
Bar Chart
Add Map
Save

Map: President Birthplaces [remove][edit]

A Map displays location-based data on a Google Map.

Visualization Title	President Birthplaces
Location field	birthLatLng contains a Lat,Lng
Popup Size	Width: 200 px, Height: 200 px
Marker Size	Width: 60 px, Height: 60 px
Only show items of type	President (Optional)
Icon	imageURL (Optional)

Fig. 4. Adding a Data Visualization

The wizard consists of four main steps:

Add Visualizations Supported visualization types include lists, tables, maps, timelines, scatter plots, and bar charts.

Add Facets Add faceted navigation to the visualization. Supported facet types include free-text search, list facets, range sliders, and tag clouds.

Configure Display Many blogs follow a narrow-width article format while some rich visualizations are wide, so DataPress includes an "lightbox" option which presents visualizations as YouTube-style previews that expand to hover over the full web page when clicked. This step in the wizard also allows the blogger to link custom CSS files to the visualization.

Lenses Lenses may be thought of as data style sheets—they are templates that define how items of a particular type should be displayed. DataPress provides a WYSIWYG lens editor that includes support for images whose URLs appear in the data.

Because DataPress is aware of the data that has already been associated with the blog post, it is able to suggest values for many of the configuration options required to create a visualization. When each new data item is added to the blog post, DataPress uses the Exhibit framework to parse the data in the background and update a running list of the item types and properties relevant to the visualization. This is particularly useful if a user is linking to data from another site on the web. Without even looking at the raw data or schema, the user is able to immediately begin crafting a visualization, with data-aware autosuggest fields providing the possible answers to necessary questions.

Once a visualization is configured, it can be inserted into the blog post by clicking a button in the wizard. The visualization appears in the blog text editor as the placeholder token {{`Exhibit`}} to mark its desired placement. Users can always re-edit their visualization by clicking on the toolbar button once again (we currently only support one visualization per blog post).

4.4 Data Sharing

After data has been associated with blog posts in DataPress, it can be shared with others in two different ways. The first is by nature of the fact that blog posts created with DataPress have links both visible (as optional data footnotes) and invisible (as links embedded in the markup of the page) that allow others to re-use the data associated with the post. Other bloggers with DataPress, for example, need only reference the URL of a data-laden blog post to automatically import all of its data and begin crafting visualizations to rebut, reinforce, or simply echo the message.

The second, and more intriguing, form of data sharing is made possible via data feeds. Just as WordPress allows RSS readers to fetch custom feeds specific to a particular tag or category of post, DataPress responds to requests to assemble data item feeds along similar lines. This feed generator creates aggregate collections of data items for a particular tag or category tracked by the blog. It does so by grouping together all data items published with posts that are marked with the specified tag or category. The following URL is an example of such a request:

```
/.../datapress/feed.php?tag=Research+Paper
```

Using data feeds web users may fetch a feed (in either JSON or RDF) of the structured data added to blog posts and incorporate that data into their own visualizations. A research group, for example, could aggregate the individual users' reading blogs into a group-wide record of readings.

If we accept that many bloggers blog out of the hope that others will consume what they blog, we can conclude that bloggers will be attracted to the idea of offering rich consumable data feeds with no additional effort on their part. We believe that such an access methodology will encourage increased *casual data curation*, as users who blog about similar items over time (trips, meals, workouts, papers, etc.) will value the data feed more than the sum of each individual annotated item once others can present the data in a visual, interactive manner.

5 DataPress in a Data Ecosystem

While the first step toward data publishing for bloggers is to give them value for using structured data, we keep in view the eventual goal of integrating linked data sources across the web. One natural link is the one between blogs run by individuals and wikis curated by communities. Projects like Semantic MediaWiki (SMW) and Freebase already offer several tools to support community-curated datasets. This section describes DataPress' features for integrating into such an ecosystem of data publishing.

To demonstrate the possibilities of such an ecosystem, we extend SMW with a plugin we have developed called Wibit[4]. Wibit enables interactive visualizations, data sharing, and schema sharing using the data contained in the SMW knowledge base. From a visualization perspective, Wibit provides a WikiText syntax that enables SMW users to create Exhibit visualizations that aggregate the results of an `ASK` query (to be contrasted with approaches like Project Halo [10] which make use of graphical interfaces). From a data perspective, our development version of Wibit provides a data API that permits external services to query the wiki knowledge base.

Working together with DataPress, the Wibit extension provides a number of integration points between SMW and data-aware blogs. Using the Wibit API, DataPress users can issue a remote `ASK` query and visualize its results from within a blog post. As DataPress allows multiple data sets to be combined, this means that a blogger can combine a wiki's data set with their own data feeds. This data flow also works in reverse: Wibit can aggregate data feeds across several blogs to display a visualization of blogged items.

As the data web evolves, we believe this blog-wiki connection is also a mechanism to encourage schema convergence within communities of interest. Users of a community can collaborate on the common definition of an item type on their community wiki, and then bloggers can use this schema to publish instances with their blog articles. Wibit's API exports SMW schemas in a JSON format for the DataPress template loader to read. When DataPress users are adding

[4] A wiki running the introduced extensions and examples is available at: `http://projects.csail.mit.edu/wibit`

data items to their posts, they may pick from one of these community-defined item schemas instead of creating their own.

By facilitating the transfer of visualizations, data and schemas across blogs and wikis, data-oriented tasks can live closest to where they are natural: wikis for crowd curation and blogs for individual publication and reflection.

6 Lessons Learned

DataPress is available as an open sourced plugin for WordPress. We now describe initial observations about how early adopters have used DataPress and provide lessons learned from phone interviews with three of these DataPress users. The users in this study are all running DataPress 1.2 or earlier, which lacks individual item publishing, data feeds, and Semantic MediaWiki hooks.

6.1 DataPress in the Wild

Since releasing DataPress, the tool has been downloaded 90 times. Of these downloads, 21 users chose to participate in a statistic collection study the software offers as an option, including one website whose DataPress-built exhibit has seen over 55,000 page views. We present some observations about this log data, though we stress that the number of users does not give our results statistical significance. In total, 56 visualizations were created and reported back to our servers, receiving 64,324 total page views. Of these, approximately half were created as permanent pages on their site, while the other half were embedded in blog entries.

Facets, or Lack Thereof. Facets are an important component for navigating structured data, and many Exhibits found online are heavily faceted to support deeper navigation of the data. One might expect highly-faceted Exhibit configurations through DataPress, but many DataPress-based Exhibits we found consisted of simply an unfaceted map or timeline for inline display of data. This suggests that even simple tools without the interactive features Exhibit provides—such as search, faceted navigation, and data lenses—are of great help to bloggers with data to display.

Lightboxing. Because of the narrow-width layout of many blogs, we assumed that the lightboxing feature of DataPress would be heavily used. However while lightboxed visualizations received more pageviews than "inline" visualizations in our data set, the lightbox setting was not frequently employed. User interviews indicated that some users might not have understood what the feature provided. Additionally, the reduced use of facets might have resulted in less space-constrained visualizations than we expected.

Data Footnotes. Finally, in our goal of exposing data for future reuse, we tried to make data footnotes simple to embed in a blog entry. While data footnotes are included by default as a textual token in any post that contains a DataPress-configured visualization, most visualization authors removed the footnotes from their entries; user interviews indicated the reasons are varied.

6.2 User Interviews

To better understand user motivations for seeking out DataPress and to learn how they used the tool, we conducted e-mail and phone interviews with three DataPress users: a scientist managing a publications list, a large website owner wishing to add dynamic features to HTML tables, and a hobbyist/entrepreneur who maintains a website for his citys local music scene. None of the three users were technical: the most experienced felt confident enough to edit CSS, but not write JavaScript.

None of the three users indicated “data blogging” as their goal, and none of them published a visualization inside a blog post. Rather, each used DataPress to place visualizations on permanent, dedicated pages of their own, more like a content management system than a blog. Time will tell whether this contradicts our claims that blogs are a natural place for structured data publishing. Much as Flash animations or audio files were once a destination of their own, while today they are casually embedded in blog posts, one might expect a similar transition for rich data objects.

Latent Data Needs. We claimed in Section 3 that the prevalence of data in natural language blog posts indicates a latent need for better data publishing tools. For the three users we spoke with, the need was not so latent: each actively sought a way to present dynamic data visualizations on their site. One had heard of Exhibit, and installed WordPress and DataPress in order to create Exhibits without having to edit HTML. The other two had actively searched for a data visualization tool over the course of several months and eventually found Exhibit. After trying unsuccessfully to integrate Exhibit into their WordPress installations, they returned to the web looking for help and found DataPress.

The fact that some users are searching for months to find data visualization tools suggests that the many APIs offering data visualization services have an untapped audience of bloggers who want these services but dont know what to do with an API. For each of these authors who found our tool, there are surely more that have yet to find a tool to help them, and still more who haven’t even thought to look for such tools because they believe data visualization to be outside the reach of bloggers.

Crossing the Structured Data Chasm. We learned that tools which provide a compelling reason for users to publish structured data can lead users to structure previously unstructured or poorly structured collections. One user said that she previously maintained a list of her publications in a MS Word document, but the ability to publish a faceted list of publications online motivated her to structure this list in a Google Spreadsheet. Another user initially maintained an HTML table to present a hand-curated collection of data, but moved this data into a separate data file so that he could provide users the ability to better explore the data.

Features without examples go unused. When we asked authors why they did not use some of DataPress’ features, a common response was that our interface did not show examples of what the result of those features would be. This is food for thought from a design perspective: even though the high-level function of these feature was often clear ("Add a map", for example, or "Add a search box") the users still wanted to see usage examples first. After seeing (or hearing) these, they decided that the feature would be useful to them in many cases. Our high-traffic user who was only publishing a dynamic table was so enthused about the other configuration options after speaking with us that he added a map, timeline, and several dynamic facets to his visualization the day after our interview with him.

We must accommodate a spectrum of data ownership philosophies. We also learned that users are well-aware of the potential perils of publishing reusable datasets and easily-replicated visualizations. While not deeply technical, all three users understood that someone could copy their dataset by linking to it, and replicate their visualization by copying their Exhibit HTML. Their reactions varied. One author felt ownership over his data and would want any reuse negotiated beforehand, though he recognized that symbiotic relationships could be built around collaborative data editing. This author took the time to modify the CSS of his site to hide Exhibits bundled data copying interface. Despite that, he was happy see his entire visualization embedded in another site as long as the site drove traffic back to his. Another author was fine with reuse of either her data or her visualization, but felt that reuse of both together would be inexcusable copying. Finally, the third author was fine with his visualizations and data being reused by others, as long as proper attribution and links were provided. As tool builders, we must remember to try to accommodate both the information sharer and the businessperson who seeks benefit in exclusivity.

Users want more data tools. The authors we spoke to also understood—and requested—features related to the wider data ecosystem, apart from the visualization capabilities. One mentioned encouraging other site owners to collaboratively maintain complimentary data sets so that they could display the data in different ways on their sites. Two of the three authors specifically requested the ability to communally maintain data on a wiki and then display visualizations of it from within their blog environment. This is significant because these authors were using a version of DataPress without this feature and were unaware that it existed.

7 Related Work

The past few years have seen a great number of projects devoted to visualizing and cataloging structured data on the web. Many Eyes [13] allows users to upload data files and create interesting data visualizations via a web interface. These visualizations are both viewable on their site and embeddable into other sites. While Many Eyes facilitates data visualization, it requires the user to step outside their authoring tool of choice (such as a blog or wiki) and use a third-party service to create and host their visualization. DataPress instead enables authoring from within the blog environment and without third-party services. Further, while Many Eyes focuses on numerical data and content modeling, we target faceted navigation [9] across semi-structured data sets. Semi-structured data opens doors to visualizations involving multiple datasets, allowing authors to build on discussions with novel contributions from new data.

Sense.us [11] is a study in visualizations which facilitate asynchronous collaboration in a *centralized* fashion. We want to modify this model by *decentralizing* the visualizations and data references, allowing collaboration to occur in the native content publishing platform(s) of the user(s).

Exhibit [12] is a client-side web framework for creating rich visualizations of data. Exhibit combines textual data files (such as RDF or JSON) with an HTML-embedded configuration file to produce interactive faceted data displays. While they needn't be programmers, Exhibit authors must be comfortable editing raw HTML and often must be familiar with data formats such as JSON. DataPress relies on Exhibit to power its visualizations, but it relieves the need to understand Exhibit's configuration syntax by providing a wizard that integrates with the blogging platform. In doing so, we aim to bring Exhibit's effective visualization capabilities to the broader class of users.

The Google Visualization API [4] enables programming-savvy webmasters to create a variety of data visualizations. As we aim to bring such visualizations into the realm of blog and wiki content, we see tools like this as potential components to incorporate into our own framework.

While the New York Times Visualization Lab [6] does not appear to use a generally-available framework for authoring displays, it deserves mention as an organization which puts a lot of effort to embed rich information displays in online content. The fact that the interactive data-driven diagrams appearing in its online edition appear to be hand-coded only underscore the need for better general-purpose visualization tools accessible to web authors.

Several projects have also risen to prominence to provide entry and cataloguing of structured data on the web. DBpedia [7] curates the structured information already present in Wikipedia taxonomies (categories) and Info Boxes. DBpedia crawls Wikipedia weekly and coerces that information into an RDF database. Semantic MediaWiki [14] is a MediaWiki extension that enables users to embed key-value annotations about a wiki topic directly in its article text. An alternative approach to DBpedia, Semantic MediaWiki integrates awareness of the inherent structure and types of data into the wiki, and thus the authorship process, itself rather than attempting to recover structure from the natural-language oriented MediaWiki database. Other tools, such as Freebase [2] and Factual [1] provide many-to-many *data* authorship environments rather than attempting to interweave structured data curation along with natural language information repositories. These projects are an interesting new class of democratized data management tools by themselves, and we see them as being another important public data source in the connected data ecosystem that is evolving.

8 Conclusion

The design of DataPress reflects a belief that a data-aware web needs tools that make grassroots authors *want* to work with data. We show the need for such tools with a study of data-oriented blog content. DataPress makes progress on this goal by fitting portions of the semantic web vision into a tool crafted specifically for the blogging workflow. DataPress provides bloggers with an easy way to create, link, and publish data while preserving many of the properties that make blogging an attractive publication platform: one-click publishing, flexible format support, easy copy and paste, and immediate results. DataPress further demonstrates a possible ecosystem of grassroots semantic web publishing in which community wikis serve to centralize ontology management while bloggers use these definitions to create feeds of data over time. We reflect on conversations with our users to better understand how this need is manifested and how to build better tools to facilitate casual use of structured data on the web.

References

1. Factual. `http://www.factual.com/`, Accessed October 13, 2009.
2. Freebase. `http://www.freebase.com/`, Accessed October 13, 2009.
3. About Wordpress. `http://wordpress.org/about/`, Accessed October 29, 2009.
4. Google visualization API. `http://code.google.com/apis/visualization/`, Accessed October 29, 2009.
5. Technorati. `http://technorati.com/`, Accessed October 29, 2009.
6. The New York Times Visualization Lab. `http://vizlab.nytimes.com/`, Accessed October 29, 2009.
7. S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of ISWC07*, 2007.
8. M. Butler, J. Gilbert, A. Seaborne, and K. Smathers. Data conversion, extraction and record linkage using XML and RDF tools in Project SIMILE. Technical report, HP Laboratories Bristol, 2004.
9. A. Elliott. Flamenco image browser: using metadata to improve image search during architectural design. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 69–70, New York, NY, USA, 2001. ACM.
10. N. S. Friedland and P. G. Allen. The Halo Pilot: Towards A digital Aristotle. Technical report, Vulcan, 2009.
11. J. Heer, F. B. Viégas, and M. Wattenberg. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1029–1038, New York, NY, USA, 2007. ACM.
12. D. Huynh, R. Miller, and D. Karger. Exhibit: Lightweight structured data publishing. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, 2007. ACM.
13. F. Viégas and M. Wattenberg. Shakespeare, god, and lonely hearts: transforming data access with many eyes. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 145–146, New York, NY, USA, 2008. ACM.
14. M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 585–594, New York, NY, USA, 2006. ACM.