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

Edward Benson, Adam Marcus, Fabian Howahl, David Karger
{eob,marcua,fabian,karger}@csail.mit.edu
MIT CSAIL

ABSTRACT

The web has dramatically enhanced people's ability to communicate ideas, knowledge, and opinions. But the authoring tools that most people understand, blogs and wikis, primarily guide users toward authoring text. In this work, we show that substantial gains in expressivity and communication would accrue if people could easily share richly structured information in meaningful visualizations. We then describe several extensions we have created for blogs and wikis that enable users to publish, share, and aggregate such structured information using the same workflows they apply to text. In particular, we aim to preserve those attributes that make blogs and wikis so effective: one-click access to the information, one-click publishing of content, natural authoring interfaces, and the ability to easily copy-and-paste information and visualizations from other sources.

Categories and Subject Descriptors: H.3 [**Information Systems**]: Information Storage and Retrieval

General Terms: Design, Human Factors, Management.

Keywords: Linked Data, Visualization, Blogs, Wikis.

1. INTRODUCTION

The Web revolutionized individuals' ability to communicate their knowledge and opinions to others. While previous tools emphasized one-to-one communication (email) or paternalistic managed information (AOL, Prodigy), the web was a medium in which anarchy could beneficially reign, allowing any person to author information accessible to multitudes. Among the many factors which made this possible, hosted blog and wiki platforms stand out: these tools provided users the ability to publish, collaborate, and remix documents with little or no knowledge of the underlying technologies that made the web work.

These grass roots web authoring tools primarily support text—free-flowing paragraphs of natural language, possibly augmented with a limited amount of rich media. In contrast, professional publications such as the New York Times often include interactive, data-driven content with faceted browsing and graphical visualizations. 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. The scientific community has increasingly turned to wikis

and blogs to share and discuss findings, and organizations such as the Science Commons are actively working to encourage the practice. The 2009 Knight Commission Report [1] forecasts a world in which journalism increasingly trends towards small, intensely local, web-based publications with budgets that are unable to support large information visualization operations but need such functionality all the same. A study we performed as part of this project finds that even casual bloggers regularly discuss items of structured data even if they do not see themselves as "data bloggers." These trends suggest that data and data visualizations *are* relevant to small-scale web publishing and that there exists a widespread latent need for WYSIWYG, one-click publishing tools that support data and rich visualization.

2. DATAPRESS AND WIBIT

In this work, we present two tools which demonstrate that web authorship tools can support, for rich structured data, the same behaviors that have made the web's text authoring tools so effective. DataPress, a tool for incorporating data in the WordPress blogging workflow, allows bloggers to author, embed, copy, and re-mix datasets and interactive visualizations into their blogs. Wibit, a tool for querying and visualizing data in the Semantic MediaWiki environment, allows wiki authors to aggregate and host data visualizations from within the WikiText environment.

In particular, our tools show that

- publishing structured data can dramatically enhance individuals' expressive power, letting them create content similar to that offered by professional publishers
- that such structured data and visualizations do not require the author to understand complex data models, and instead can be authored using same tools and metaphors with which grass-roots authors are already familiar: WYSIWYG editing for bloggers, and Wiki-Text editing for wiki users
- that these tools can offer the same copy-and-paste workflow as text, making it easy for authors to "quote" both the data and the visualizations authored by other users, either to be used unchanged or as a starting point for a re-mixed version

DataPress and Wibit enable cross-site data sharing and linking, allowing the blogging and wiki communities to add structured data and interactive data visualizations to the interlinked natural language conversation already present across the blog and wiki landscape. We leverage the collaborative curation environment of wikis to host community

repositories of data models to allow easy schema re-use and to encourage schema convergence where possible. Our plug-ins then provide features that allow bloggers and wiki editors to publish data (either as individual items or entire sets), aggregate data (across a wiki or a blog category, for example), and easily create faceted data visualizations.

In short, to those coming from a scientific or data-oriented background, our tools present a model of what an of ecosystem of data-centric web publication might look like – one that includes all of the characteristics that have allowed the current ecosystem of text-publishing to thrive. To those coming from a journalistic background, we believe our tools offer the significant benefit of increased expressive power and content reusability to grass-roots authors with almost no cost in changing their mental model of their content, or in changing the workflow needed to produce that content.

3. DEMONSTRATION

Consider a university department in which a graduate research group records its paper-reading and conference-going activities on a shared wiki. With todays publishing tools, such content is merely text: it is browsed in narrative form, is difficult to aggregate, and can only be shared by copying and re-entering the data elsewhere. But research papers and conferences are *not* just text—they are rich objects that can be plotted on a map or timeline, aggregated from disparate data sources, and shared by reference.

With DataPress and Wibit, this common data curation task becomes much more fruitful. Figure 1 shows a wiki page from an MIT CSAIL reading group using these tools. Here, papers entered into Semantic MediaWiki as individual pages appear aggregated on a timeline using Wibit. Wibit leverages the Exhibit[2] framework to provide faceted browsing and interactive displays of the data from the Semantic MediaWiki database—academic papers, in this case.
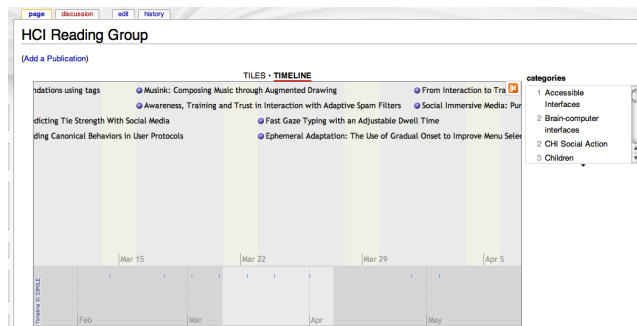

Figure 1: Wiki Timeline for HCI Reading Group

Bloggers using DataPress may also access the information stored on this, or any other Wibit-enabled wiki. The DataPress plugin provides buttons on the post editing toolbar, shown in Figure 2. These buttons allow the user to enter data, link to data (on Google Spreadsheets, or a wiki, for example), or insert a new visualization into the blog post.

Figure 2: DataPress Entry Points

DataPress and Wibit interact with users by extending the interface of their respective environments of web forms and WikiText. Adding a data visualization to a blog post with DataPress is a form-driven process, depicted in Figure 3, which guides the user through questions about the data's source, the visualization type, and the desired facets for interacting with the data. Like Wibit, DataPress uses Exhibit to create visualizations.
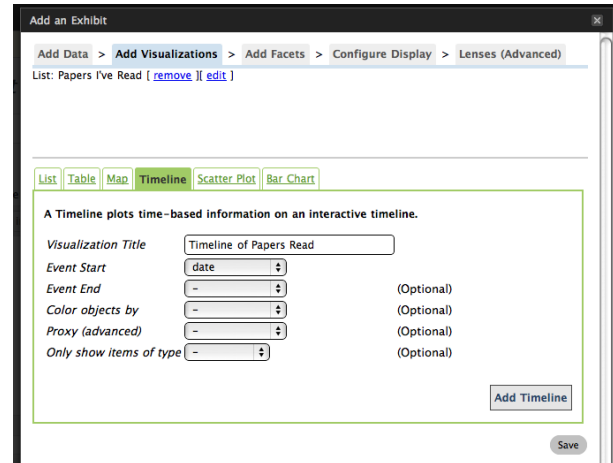

Figure 3: Blogging a Data Visualization

The final visualization, shown in Figure 4, shows a list of papers read by the HCI reading group injected into the body of the blog post. This list can be styled to the blogger's liking, and the data is fetched "live" from Wibit's aggregation feed with each page request.
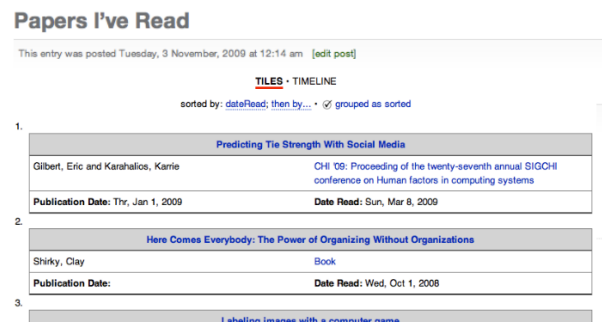

Figure 4: Wiki Data on a Blog

The design of DataPress and Wibit both reflect a belief that a data-aware web needs tools that make people *want* to work with data. With the tools we have developed, authors can take more advantage of the structure in their data without changing the way they work.

4. REFERENCES

[1] T. K. Commission. Informing communities: Sustaining democracy in the digital age. Technical report, The Aspen Institute, 2009.

[2] 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.