6. Conclusion

In this thesis I set out to explore user interfaces for supporting data-centric interactions on today's Web. In doing so, I have shown that such data-centric interactions are useful to and feasible for casual users, and usable tools can be built to support such interactions by gearing for small, simple data sets and common casual needs.

This thesis makes the following contributions:

- First, this thesis combines a simple graph-based data model with simple extensions of the HTML syntax in order to let casual users—those without programming skills—publish web pages offering advanced browsing features and rich visualizations.
- Second, this thesis shows that the cost of using web data extraction technologies can be lowered for casual users by retaining presentation elements from the original web pages. These elements preserve visual context and visual semantics so that direct manipulation techniques can be applied to augment the original pages with more features without requiring users to label fields in the extracted data.
- Third, this thesis proposes that for simple data sets, direct manipulation techniques can be used to let casual users integrate data and construct rich visualizations from it without any need for programming.
- Finally, by letting casual users publish data to the Web in browsable form, extract and reuse data from the Web, and integrate data from several sources without any need for programming, this thesis demonstrates that data-centric interactions can be offered to casual users on today's Web without having to wait for a semantic web.

This chapter continues with possible future directions from this line of investigation and concludes with a discussion on larger contexts surrounding my research.

6.1 Future Work

This section proposes future work items organized by the three core aspects of the thesis. At the end, it draws together all three aspects and proposes a way to build a data-centric browser for interacting with the future data-centric Web.

6.1.1 Publishing Data

Small-time publishers already benefit just to be able to publish 1,000 items using Exhibit. However, over time, they might develop sophistication in their publishing needs, wanting to go beyond that limit. A smooth transition path from 1,000 items to tens of thousands of items is required to migrate this class of users.

Small-time publishers will also demand easier editing interfaces. While HTML was good enough to bootstrap the Web, it was not until the arrival of blogs and wikis that user content exploded. Various user interface ideas in Potluck can be moved into Exhibit so that publishers can edit their data and their exhibit layouts directly within the browser using direct manipulations.

Users of exhibits will also want more control over Exhibit. If an exhibit is not configured in the way that best serves a user, such as not showing a map view, the user will benefit from being able to reconfigure the exhibit in-place, adding a map view. Again, Potluck's direct manipulation techniques can be adapted for this purpose.

Another future direction can explore the benefits of embedding Exhibit in online scientific publications. Currently, scientific papers are published as PDF files and the data graphics they contain are just static images. If these papers were instead published as web pages that embed Exhibit, their data graphics can be interactive and their data can easily be reused by their readers. Some technologies are needed to migrate the class of scientific publishing users from their conventional PDF publishing process over to data-centric web publishing.

If the use of Exhibit proliferates, and there are millions of data-rich web pages created by casual users, search engines should be adapted to leverage the data within these pages, providing search capabilities that go beyond keyword queries.

6.1.2 Extracting Data

Sifter lowers the cost to using web data extraction technologies by retaining much of the original presentation. But more can still be retained. For example, if used

prices are shown in a gray, stroke out typeface on the original web page, the browsing control box for used prices should also use that same typeface to display its facet values. This will strengthen the connection between a browsing control box the corresponding field values in the items.

The augmentation interface can be moved earlier in the process. For example, even before extraction, Sifter can let the user right-click on any part of the web page and invoke a sorting or filtering command, which would then trigger the extraction process and show the browsing control box for the corresponding field. The extraction over subsequent pages can also be done incrementally. That is, the augmentation interface can be shown right away and the extraction process updates it gradually as more and more pages are scraped. This change in the workflow lets the user invoke the desired feature earlier and start to get the results earlier.

Sifter should also let the user cancel the incremental extraction process or make correction at any point in time. Correction may involve, for example, splitting a field into several fields, merging several fields into one, and changing the value type of a field.

6.1.3 Integrating Data

The idea of retaining original presentation can be carried over from Sifter into Potluck so that the user will not be faced with raw data at the beginning. Research is needed on how to compose the presentations from different sources. For example, if one original site offers only a map view, and another offers only a timeline view, what should Potluck show at the beginning? While an obvious answer would be to juxtapose the views, as the views remain separate, this solution might actually discourage the user from trying to mix the data together.

Once a user has integrated data from some sources and constructed visualizations for that data, the actions she took to align the data, to fix it up syntactically, and to construct the visualizations can be stored and used to make suggestions to other users who want to deal with the same sources.

6.1.4 Toward a Data-Centric Browser

As there is more and more data in reusable forms on the Web, casual users will encounter it more often and their use of the data will increase in frequency and in sophistication. Where casual users meet the Web—the web browser—will need to catch up. Designed for the text-centric Web for viewing hypertext documents, the contemporary browser may no longer be suitable for the data-centric Web. A data-centric browser is needed to address casual users' needs in interacting with a future data-centric Web.

One of the Web's greatest strengths is in its generality. Documents on any topic can be published to the Web and viewed and interacted with in any web browser. Each web browser provides a set of basic, generic features that are sufficiently useful for browsing anywhere on the Web. Similarly, every data-centric web browser

should provide some set of basic, generic features sufficiently useful for interacting with any data encountered on the data-centric Web.

The chief advantage of a data-centric Web over a text-centric Web is the possibility of repurposing data—using it in ways different from how it was originally intended to be used. In order to repurpose data you must first be able to *retrieve* it, which might involve screen scraping for web pages that have not become data-centric. Then, as shown in the previous two chapters, repurposing data involves *browsing* through it in ways not supported at the original sites, *merging* data from several sites together, *editing* the data to clean it up and unify it syntactically, and *visualizing* it in novel ways.

Repurposing data yields new data that is worth *saving*. Even original data can be worth saving, just like some web pages are worth bookmarking to some users. Once there is support for saving data, there is a need to interact with the saved data in the long term. This involves *browsing* and *searching* through the data later, which might warrant *organizing* it earlier on.

Finally, an ultimate goal for data collecting and repurposing might be *publishing* it back into the Web.

Thus, a data-centric browser should support the following features:

- retrieve data from any web source;
- merge data from several sources;
- edit data efficiently;
- visualize data in many ways;
- browse/search through data;
- save data permanently;
- organize data;
- publish data expressively.

The various components in this thesis provide answers to how these features can be supported. More research is needed to fit them all together into a coherent whole and provide a seamless experience for moving data out of and into the Web.

6.2 Discussion

Now at the end of the thesis, let me venture outside the realm of computer science into the larger social, political, and economic contexts surrounding the idea of a data-centric Web.

6.2.1 Data Publishing as Empowerment

The diversity of interests on the Web brings big profits to online retailers. On Amazon, there are hundreds of thousands of unpopular books, each selling only a few to a few hundred copies, whose combined profit rivals that of the few hundred

books that sell millions of copies. This observation is explored in "The Long Tail" [39].

If we rank the various kinds of information that people want to publish on the Web by their popularity, we should also get a long tail (Figure 6.1): a few dozens of readily recalled kinds of information such as consumer products, news articles, events, locations, photos, videos, music, and software projects populate the massive head of the curve while hundreds of thousands of unpopular topics, such as sugar package collection and lock picking, spread thinly over the long tail. Clustering search engines such as clusty.com show dozens of sites of each unpopular topic. Wikipedia's several thousand categories and DMoz's half a million categories show how long the tail is.

Blogs have made publishing text articles even easier than writing HTML documents from scratch. Wikis let small to large communities collaboratively grow and maintain reasonably-sized web sites of text articles. Multimedia publishing services like Flickr and YouTube, complemented by the arrival of affordable consumer cameras, also add to the publishing power of the mass population. These publishing venues are now widely adopted and highly appreciated for giving the mass population text-based and multimedia publishing power rivaling large, professional, well-funded organizations. Wikipedia has grown large enough to be compared against Encyclopedia Britannica [48]. "Citizen journalism" [2] can deliver news faster and from more angles than news networks.

As blogs, wikis, and other publishing mechanisms spread, they become commonplace. A well-styled web site is now the baseline—the least to be expected. It sparks no surprise nor delight, and does not carry any authority just by its look. To set itself apart, a web site must differentiate against other sites by other factors, such as the expensive features that it can afford or the perceived objectivity in which it communicates its view. For example, to report a presidential election, a web site might show a map plotting per-state support for each political party, and a timeline coupled with the map to show how such support has changed over time. The site might even make accessible the data that fuels the map and the timeline so that

Figure 6.1. The Long Tail of Information Domains.

its visitors can double-check the validity of its visualizations. In doing so, the site builds its public image of being objective and open.

In my research, Sifter and Potluck, which allow casual users to repurpose and mix data from the Web, can be seen as anticipation of such a future of data-rich web sites. To complement, Exhibit gives the mass population ammunition to compete against large content providers as the two sides co-evolve toward that future.

6.2.2 The Data Market

The motivations for publishing data mentioned in the previous section are based on competition. Publishing data accessorizes one's web site and sets it apart from the rest. Reputation can be accumulated if one's data is reused for some grand purposes or by some reputable parties. If quality of data is a competitive advantage—a reasonable assumption—then competition among publishers will drive data quality upward. Perhaps economic theories can be applied to set the desirable conditions that drive this *data market* to produce the desired data in the end.

If there will indeed be a market in which data is traded for reputation or sold for money as small-time publishers try to differentiate among themselves, new industries will spring up to support their needs. Software and services for collecting, editing, managing, and hosting data are in demand. Data needs to be verified and certified by independent parties. Licenses are required to facilitate reuse. The right tools and services produced by these industries can help develop the market.

There is already some infrastructure for and evidence of such a market. For example, online consumer-oriented database and spreadsheet services like DabbleDB [5] let people collect and share data. Many Eyes [16] and Swivel [35] let people construct data visualizations. Sense.us [51] lets people collaboratively scrutinize data visualizations, pointing out anomalies and explaining trends. The Science Commons [29] are working on licenses for scientific data.

6.2.3 Data as Common Goods

Building such a data-rich Web has long been advocated by the Semantic Web project. Semantic Web researchers initially appealed to the vision in which software agents could harvest information on the Web and perform information-centric tasks on behalf of people. For that vision to work, the Semantic Web must be sufficiently complete and coherent such that software agents have adequate and reliable data to work on. This strategy requires people to believe in the vision and then to cooperate to build the Semantic Web for the benefits of all humanity in the future. As there is no immediate benefit but much initial labor required, this strategy mostly appeals to and brings together far-thinking, altruistic individuals. However, if the labor required is small enough and the vision very appealing, this strategy might even appeal to the mass population, as in many cases where *crowdsourcing* [3] does work. For example, there have been efforts by Google and OpenStreetMap to let web users edit street maps for countries in which street maps cannot be obtained

from the government. FreeBase [8] intends to let web users build a global database equivalent to Wikipedia.

We can also imagine smaller efforts in which geographically small communities collaborate to maintain databases of their resources for emergency purposes. Semantic MediaWiki [72] might be a good starting point toward this end. When a natural disaster strikes, it is desirable to get a big picture of the situation as quickly as possible. How many adults, children, elders, and physically disabled are in the disaster zone? How many vehicles are available? Does anyone own a boat if the disaster is a flood? Where do the doctors and nurses live? Where do mechanical engineers and electricians live? Which buildings are most likely still standing and capable of housing a certain number of people? Survivors can even help to collect data on the fly wherever they are. People outside the disaster zone can help organize and analyze the collected data.

Even before disasters strike, these community-maintained databases can be anonymized and pooled together to give a big picture of the whole country. Through incentive schemes, the population can be redistributed so that each region within the country can be as autonomous as possible, having all the skills and resources needed for disaster response.

6.2.4 Data as Culture

To engage the mass population for the purpose of building a data Web, usable technologies may not be enough. Various media machineries might need to be called upon to spread the meme of data appreciation to every corner of society.

Television shows such as "CSI: Crime Scene Investigation", which show unrealistic scientific capabilities and romanticize crime scene evidence in rich visualizations, have raised real-world expectations of forensic science [4]. More students are enrolling in forensic science programs recently. A similar means can be used to increase the public's expectation for the availability of data that helps during crises. Such expectation might in turn increase the public's willingness to contribute to the community-maintained databases suggested previously.

The seed of appreciation for data can also be sown early in grade school education. Young children can be given tools and instructions for experimenting with data so to develop skills and tastes for data. Just as science kits are sold to pique children's interests in science and give them hands-on experience with scientific experiments at an early age, "data kits" can also be made available to children.

What would it be like to live and grow up in a data-rich society? Would it be different from the *information*-rich society in which we are currently living? We are already facing the information overload problem. Would it not be worse to be overloaded with raw data? These questions remain in the realm of media studies, beyond the scope of this thesis.