User Interfaces Supporting
Casual Data-Centric Interactions on the Web

by

David F. Huynh

S.M. Computer Science and Engineering, Massachusetts Institute of Technology (2003)
B.A.Sc. Computer Engineering, University of Waterloo (2001)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2007

© Massachusetts Institute of Technology 2007. All rights reserved.
User Interfaces Supporting
Casual Data-Centric Interactions on the Web

by

David F. Huynh

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2007, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

Today’s Web is full of structured data, but much of it is transmitted in natural language
text or binary images that are not conducive to further machine processing by the time it
reaches the user’s web browser. Consequently, casual users—those without programming
skills—are limited to whatever features that web sites offer. Encountering a few dozens of
addresses of public schools listed in a table on one web site and a few dozens of private
schools on another web site, a casual user would have to painstakingly copy and paste
each and every address into an online map service, copy and paste the schools’ names, to
get a unified view of where the schools are relative to her home. Any more sophisticated
operations on data encountered on the Web—such as re-plotting the results of a scientific
experiment found online just because the user wants to test a different theory—would be
tremendously difficult.

Conversely, to publish structured data to the Web, a casual user settles for static data
files or HTML pages that offer none of the features provided by commercial sites such as
searching, filtering, maps, timelines, etc., or even as basic a feature as sorting. To offer a rich
experience on her site, the casual user must single-handedly build a three-tier web applica-
tion that normally takes a team of engineers several months.

This thesis explores user interfaces for casual users—those without programming
skills—to extract and reuse data from today’s Web as well as publish data into the Web in
richly browsable and reusable form. By assuming that casual users most often deal with
small and simple data sets, declarative syntaxes and direct manipulation techniques can be
supported for tasks previously done only with programming in experts’ tools.

User studies indicated that tools built with such declarative syntaxes and direct ma-
nipulation techniques could be used by casual users. Moreover, the data publishing tool
built from this research has been used by actual users on the Web for many purposes, from
presenting educational materials in classroom to listing products for very small businesses.

Thesis Supervisors: David R. Karger and Robert C. Miller
Titles: Professors of Computer Science and Engineering
to my parents
who are my constant source of courage
Acknowledgements

I would like to thank my thesis advisors, David R. Karger and Robert C. Miller, for guiding me when I was lost and for trusting me to explore freely on my own when I found my direction. I could not have had better supervision for every step of this arduous journey, or more gratification and pride at the end.

The SIMILE project members have also been instrumental. I thank Stefano Mazzocchi, Ryan Lee, Andrew Plotkin, Ben Hyde, Richard Rodgers, V. Alex Brennen, Eric Miller, and MacKenzie Smith for their continuing encouragement and support in various forms as well as their perspectives, insights, and wisdom that bring practicality to my research. Without SIMILE, my research would not have had the same impact and reach.

The Haystack group and User Interface Design group members have provided tremendously insightful feedback on my research, and many of them have been my travel companions throughout this long journey. Many thanks go to Vineet Sinha, Jaime Teevan, Karun Bakshi, Nick Matsakis, Harr Chen, Yuan Shen, Michael Bernstein, Adam Marcus, Sacha Zyto, Max Goldman, Greg Little, and Jones Yu.

I have also enjoyed many exchanges with users of my software published through SIMILE. I thank Johan Sundström, Josh Aresty, and Keith Alexander for their thoughts as well as their code patches, not to mention their enthusiasm in singing to their friends more praises of my work than it deserves. There can be no more satisfaction to a tool builder than to see his tools used for real.

Many people have participated in my user studies. I am grateful for their time and their insightful feedback. I also would like to thank those who have helped recruit subjects for my studies: William Reilly, Robert Wolfe, Ann Whiteside, and Rebecca Lubas.

Over the years, I have enjoyed stimulating conversations with many other people regarding my research. I thank Paolo Cicaresce, Steven Drucker, Mira Dontcheva, Eric Neumann, Daniel Tunkelang, Kingsley Idehen, Ivan Herman, Michael Bergman, Jon Crump, Danny Ayers, Vinay Mohta, Ian Jacobs, Sandro Hawke, Wing Yung, Lee Feigenbaum, Ben Szekely, Emmanuel Pietriga, Chris Bizer, Justin Boyan, Glen McDonald, Michael Bolin, Alex Faaborg, Ora Lassila, Deepali Khushraj, Ralph Swick, Daniel Weitzner, Ann Bassetti, Chao Wang, Gregory Marton, Boris Katz, and Steve Garland.

continued on next page
My friends have made the journey much more fun. I thank Sayan Mitra, Vineet Sinha, Han-Pang Chiu, Harold Fox, Agnes Ng, Rui Fan, Karun Bakshi, Stefano Mazzocchi, Ryan Lee, Paolo Ciccarese, Heidi Pan, Max van Kleek, Gregory Marton, Michael Bernstein, Olya Veselova, Brooke Cowan, and Jimmy Lin. Special thanks go to 李真 for keeping my company everyday throughout.

Whereas my colleagues asked many deep questions, my family asked a very simple one, “Are you done yet?” It was deceptively simple: while I could often provide clear responses to my colleagues, I always had to mumble some answer to my family. But now the reply is a resounding “Yes.” I’ve come to the end of my seemingly quixotic journey, and I’m grateful to my family for being so caring and anxious to always ask, “Are you done yet?” but also for being so understanding to never question why.
Whatever you do will be insignificant, but it is very important that you do it.

— Mahatma Gandhi —
CONTENTS

1. INTRODUCTION ............................................................. 17
   1.1 The Case for Data-Centric Interactions ......................... 19
   1.2 Approach ............................................................... 21
       1.2.1 Assumptions and Limitations 21
       1.2.2 Publishing Data 22
       1.2.3 Extracting Data 26
       1.2.4 Integrating Data 28
   1.3 Implications for Semantic Web Research ...................... 30
   1.4 Contributions ....................................................... 31
   1.5 Thesis Outline ..................................................... 32

2. RELATED WORK ............................................................ 33
   2.1 User Interfaces for Browsing ..................................... 33
   2.2 Publishing Data ..................................................... 34
       2.2.1 Rationales for Using Data Technologies in Publishing 34
       2.2.2 Flexibility of Data Modeling and Presentation 36
       2.2.3 Costs of Using Data Technologies 38
       2.2.4 Client-side Data Technologies 38
       2.2.5 Presentation Templating Technologies 39
   2.3 Extracting Data ..................................................... 40
       2.3.1 Goals for Web Data Extraction 40
       2.3.2 Supervised vs. Unsupervised Extraction 41
       2.3.3 Web Content Augmentation 42
       2.3.4 Facilitating Web Data Extraction 42
2.4 **Integrating Data** ................................................................. 42
   2.4.1 Data Integration and Warehousing 43
   2.4.2 Ontology Alignment 43
   2.4.3 Data Integration User Interfaces 44
2.5 **Toward a Data-Centric Browser** ........................................... 45
   2.5.1 Data Models 45
   2.5.2 Data-Centric Features 45
   2.5.3 Deployment Paths 46

3. **Publishing Data** ................................................................. 47
   3.1 **Interface Design** ............................................................. 49
      3.1.1 User Interface 49
      3.1.2 Publisher Interface 52
   3.2 **Data Model** ................................................................. 55
      3.2.1 Items 55
      3.2.2 Types 57
      3.2.3 Properties 58
      3.2.4 Expressions 58
      3.2.5 Data Import/Export 61
   3.3 **User Interface Model** ...................................................... 62
      3.3.1 Collections 62
      3.3.2 Facets 62
      3.3.3 Views 63
      3.3.4 Lenses 63
      3.3.5 UI Contexts and Formatting Rules 64
      3.3.6 Coders 67
   3.4 **Implementation** ........................................................... 68
   3.5 **Evaluation** ................................................................. 70
      3.5.1 Performance 70
      3.5.2 Usage 75
   3.6 **Summary** .................................................................. 76

4. **Extracting Data** ................................................................. 89
   4.1 **User Interface Design** ...................................................... 90
      4.1.1 Extraction User Interface 90
      4.1.2 Augmentation User Interface 92
   4.2 **Data Extraction** ............................................................. 95
      4.2.1 Item Detection 95
      4.2.2 Subsequent-Page Detection 97
      4.2.3 Field Detection 99
4.3  Evaluation .................................................. 100
   4.3.1  Evaluation of Data Extraction  100
   4.3.2  Evaluation of User Interface  100
4.4  Summary ................................................. 106

5.  INTEGRATING DATA ........................................ 107
   5.1  Scenario .................................................. 108
   5.2  User Interface ......................................... 109
      5.2.1  Creating columns and facets  112
      5.2.2  Merging fields  113
      5.2.3  Simultaneous editing  114
      5.2.4  Faceted Browsing  115
      5.2.5  Visualizations  116
      5.2.6  Miscellany  116
   5.3  Implementation ........................................ 116
   5.4  Evaluation ............................................. 118
      5.4.1  Design and Procedure  118
      5.4.2  Participants and Apparatus  119
      5.4.3  Results  119
   5.5  Summary ................................................. 121

6.  CONCLUSION .................................................. 123
   6.1  Future Work ............................................. 124
      6.1.1  Publishing Data  124
      6.1.2  Extracting Data  124
      6.1.3  Integrating Data  125
      6.1.4  Toward a Data-Centric Browser  125
   6.2  Discussion ............................................. 126
      6.2.1  Data Publishing as Empowerment  126
      6.2.2  The Data Market  128
      6.2.3  Data as Common Goods  128
      6.2.4  Data as Culture  129

7.  BIBLIOGRAPHY .............................................. 131