Experiences with Content Extraction from the Web

Mira Dontcheva[1,2] **Sharon Lin**[1] **Steven M. Drucker**[3] **David Salesin**[1,2] **Michael F. Cohen**[4]

[1]Computer Science & Engineering
University of Washington
Seattle, WA 98105-4615
{mirad, sdlin}@cs.washington.edu

[2]Adobe Systems
801 N. 34th Street
Seattle, WA 98103
salesin@adobe.com

[3]Microsoft LiveLabs, [4]Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
{sdrucker, mcohen}@microsoft.com

ABSTRACT

We present the results of a ten-week field study that explored the use of automatic Web tools for collecting and organizing Web content in the context of users' personal tasks. Our findings show that people welcome automatic gathering of structured information, such as job or rental listings, and are eager to use rich visualizations and displays of content they find on the Web. We also found that users collect a variety of Web content including a large amount of unstructured information and are interested in using automation not just for long-term content intensive tasks but also for short-lived transient tasks. Finally, we present a first exploration of an online collaborative repository of user-defined semantic content. Our study participants used this repository and modified the collaborative content to accomplish tasks.

ACM Classification H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General Terms Design, Human Factors

Author Keywords

extraction patterns, layout templates, collaboration, case study.

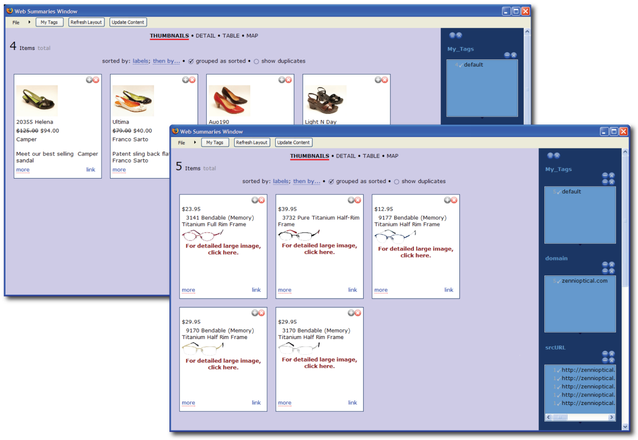
INTRODUCTION

As more and more information becomes available on the Web, the tools for collecting, organizing, and sharing Web content are becoming increasingly sophisticated. Researchers have demonstrated a variety of tools for semi-automatically gathering Web content [5] [1] [8] [16] [6], re-presenting webpages for different purposes and devices [15] [7] [13], and customizing search interfaces to better suit user needs [10] [4]. The Open Source community has enabled user-contributed automation through browser extensions. With the GreaseMonkey Firefox extension developers can quickly and easily customize any website or modify the browser behavior. The vision of the Semantic Web [3] promises to create even more sophisticated Web tools including agents that can extract, aggregate, and analyze information from many disparate sources. In this paper we explore a collaborative repository of user-defined semantic content in the context of a ten-week field study. If popularized, such a collaborative repository could help create a user-defined Semantic Web.

Previously [5], we developed the Web Summaries system, which allows users to collect and organize Web content semi-automatically. With Web Summaries the user interactively clips and tags pieces of webpages, thereby creating extraction patterns that can be applied to collect more content from similar pages. *Extraction patterns* specify the locations of selected elements using the structure of the webpage. When the user visits a new page, he can automatically collect all of the content that is analogous to the content previously selected as relevant. To organize and present the user's clippings, Web Summaries employs layout templates that create visual summaries. A *layout template* specifies how the content should be organized and formatted and defines a set of user interactions available within a summary. The templates filter and present the content according to the tags specified by the user during the clipping process.

In this paper, we extend the Web Summaries framework to include a collaborative online community and present the results of a ten-week field study in which we deployed Web Summaries to 24 participants. Although there are a number of semi-automatic tools for collecting and organizing Web content, few of these automatic tools have been evaluated in the field with users' personal tasks. We conducted this first longitudinal study of semi-automatic tools for collecting Web content with several goals in mind. First, we wanted to evaluate the utility of automatic extraction and understand when and how often users are in situations where automatic Web content extraction is useful. Second, we wanted to gain a better understanding for how people collect and organize information. Although there have been a number of ethnographic studies [14] [2] [12] exploring user behavior patterns when dealing with information, they do not reveal the granularity or type of information users collect during exploratory Web research. Finally, we wanted to explore how a community of users can benefit one another through an online repository of shared semantic content.

Our study revealed that users collect a variety of Web content including highly structured content, such as tables and lists, and highly unstructured content, such as entire articles. Our participants actively used automatic retrieval on many websites and for their own personal tasks. They created over 250 extraction patterns by clipping pieces of webpages and collected over 1000 items automatically. While Web Summaries was useful for content intensive tasks, users found it less applicable for transient daily tasks, as the tool required opening a window and actively saving content. Many participants also used Web Summaries as a mechanism for storing permanent versions of Web content that they view as highly dynamic. The participants were very positive about the layout templates and the resulting summaries. They used a variety of layout templates and requested more flexible cus-



20355 Helena
~~$125.00~~ $94.00
Camper
Meet our best selling Camper sandal

Ultima
~~$79.00~~ $40.00
Franco Sarto
Patent sling back fla
Franco Sarto

Auo190

Light N Day

$23.95
3141 Bendable (Memory) Titanium Full Rim Frame
For detailed large image, click here.

$39.95
3732 Pure Titanium Half-Rim Frame
For detailed large image, click here.

$12.95
9177 Bendable (Memory) Titanium Half Rim Frame
For detailed large image, click here.

$29.95
9170 Bendable (Memory) Titanium Half Rim Frame
For detailed large image, click here.

$29.95
3170 Bendable (Memory) Titanium Half Rim Frame
For detailed large image, click here.

Figure 1. The participants used automatic content gathering for a variety of shopping tasks, including searching for shoes and glasses.

tomizable templates. Finally, the participants used the community pattern repository to download patterns created by others. They often modified downloaded patterns to suit their own needs. They suggested many improvements to the interface to help make sharing patterns as fast and easy as creating them through clippings.

Our findings lead us to the following design implications for future semi-automatic and automatic tools for collecting Web content. First, tools for collecting and organizing Web content need to support the ability to collect both structured and unstructured Web content and display a heterogenous collection of content. Second, in order to become well integrated into a user's daily Web usage, Semantic Web tools need to be sensitive not only to long and permanent Web tasks but also to transient and short-lived tasks, otherwise users will not integrate them into their Web browsing habits. Since Web browsing is such an integral part of user's lives, tools that support information management must be fluidly integrated into the browser and be available at any time. Finally, an online repository of semantic information is still very new to users, and visualizations for exposing the available information must be carefully designed such that the semantic information can aid rather than hinder users from accomplishing their tasks.

THE SYSTEM

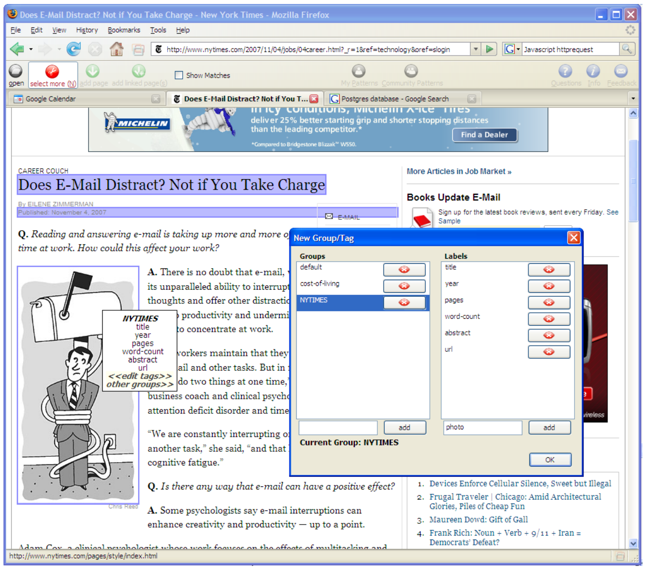
Web Summaries is implemented as an extension to the Firefox browser and is presented to the user through a toolbar. The toolbar opens a window where the user can collect Web content and also includes buttons for creating extraction patterns and saving content. The entire system is implemented as a browser extension and is written in Javascript and XUL.

In order to release Web Summaries in the field, several modification of the original implementation were necessary. Many users requested filtering and sorting capabilities for their collections of content, and we added these operations to Web Summaries by creating layout templates using Exhibit [9]. Exhibit is a lightweight Javascript framework for publishing structured data. It combines structured JSON files that describe data with presentation parameters to produce webpages that allow users to organize and filter data interactively. We implemented four layout templates – a thumbnail view (Figure 1), a detail (Figure 12) view, a table (Figure 14), and a map (Figure 9). The blue pane on the right of the display (see Figure 1) allowed users to tag their collection items or filter the collection using a facet, such as domain, review, or price. Users could also sort the collection using the facets. For example, sorting according to name sorts the items in alphabetical order. Users could also at any time press the "update content" button at the top and dynamically update all of the content in the summary. This allowed them to retrieve the most recent content from any webpage and also add new content from pages they had already visited.

In addition to creating layout templates that include filtering and sorting, we also added a collaborative component to Web Summaries through a *community pattern repository*. The community pattern repository stores all patterns created by the study participants, thereby allowing them to share patterns with one another and among devices. As already described, a pattern can tag and extract content from a webpage. Thus, this community pattern repository holds a semantic description for all of the webpages and websites visited by the subjects. Each user has a *personal pattern repository* that includes only previously created and used patterns. The union of all personal pattern repositories creates the community pattern repository. The community pattern repository was implemented with a PostgreSQL database and an Apache Web server. All communication between the extension and the server used the Javascript HTTPRequest object. All server scripts for accessing the database were written in PHP.

The interface

Figures 2 and 3 show the toolbar interface for the application that we deployed in the field study. The "open" button ● starts Web Summaries, opens the summaries window, and enables the toolbar. Since only one summary window can be open for each browser window, subsequent presses of the "open" button bring the summary window in front of any other windows. Clicking on the modal "select new" button ● enables selection and tagging of webpage elements. Figure 2 shows the selection and tagging of an image. The user can tag webpage elements with the default tags or he can create his own tags and categories of tags. In Figure 2 the user has created a NYTIMES category and is adding a 'photo' tag. The tags determine how the element is displayed in the summary. For example, an "address" tag tells the system that the clipped content can be displayed on a map. To finish selecting and tagging page elements users click the "select new" button again. The extension adds the selected content to the summary and creates an extraction pattern for that page using the selected elements. The "select more" button ● allows users to add more elements to existing extraction patterns. The "select new" button becomes a "select more" button if the user has created an extraction pattern for the type of webpage currently viewed. Thus as the user browses the Web this button changes from ● to ● as needed. The "add page" button ⊕ allows users to automaticaly add content to their summary with an extraction pattern.


Does E-Mail Distract? Not if You Take Charge - New York Times - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.nytimes.com/2007/11/04/jobs/04career.html?_r=1&ref=technology&oref=slogin
open
select more (N)
add page
add linked page(s)
Show Matches
My Patterns
Community Patterns
Questions
Info
Feedback
Google Calendar
Does E-Mail Distract? Not if You T...
Postgres database - Google Search
MICHELIN
Find a Dealer
E-MAIL
http://www.nytimes.com/pages/style/index.html

CAREER COUCH

Does E-Mail Distract? Not if You Take Charge

By EILENE ZIMMERMAN
Published: November 4, 2007

Q. *Reading and answering e-mail is taking up more and more of time at work. How could this affect your work?*

NYTIMES
title
year
pages
word-count
abstract
url
<<edit tags>>
other groups>>

More Articles in Job Market »

Books Update E-Mail

Sign up for the latest book reviews, sent every Friday. See Sample

New Group/Tag

Groups	Labels
default	title
cost-of-living	year
NYTIMES	pages
	word-count
	abstract
	url

add photo add

Current Group: NYTIMES

OK

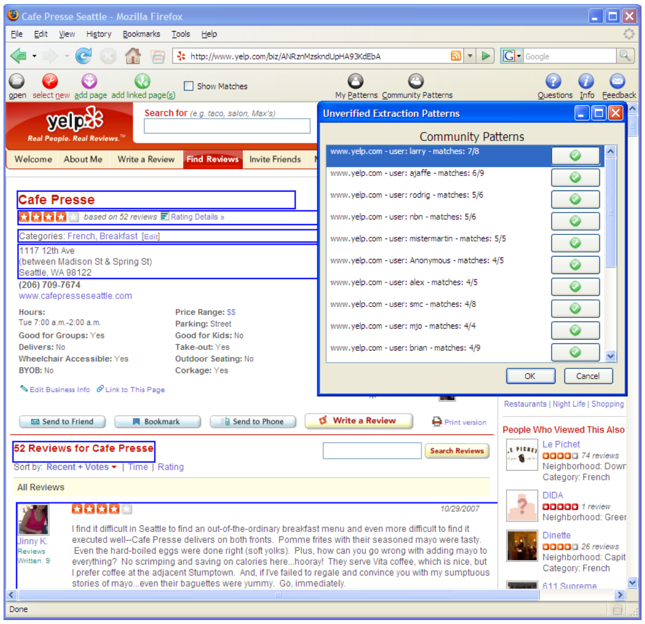
1. Devices Enforce Cellular Silence, Sweet but Illegal
2. Frugal Traveler | Chicago: Amid Architectural Glories, Piles of Cheap Fun
3. Maureen Dowd: Gift of Gall
4. Frank Rich: Noun + Verb + 9/11 + Iran = Democrats' Defeat?

Figure 2. The user clips and tags pieces of a webpage. He can create new tags and categories at any time.

This button is only enabled if there is an appropriate extraction pattern for the current page. If an extraction pattern for the current page exists in the community pattern repository, the add page button changes color and icon. The user can click on the button to see a list of matching patterns. Figure 3 shows the interface for viewing the shared database of patterns. We discuss this interface in more detail in the next section. The "add linked pages" button allows users to select hyperlinks to pages they want to add to their summary. Web Summaries follows each hyperlink, extracts the content from the hyperlinked page and adds the extracted content to the summary. The "show matches" checkbox lets users see all of the matching elements before saving them in the summary. When this option is checked, all elements in a loaded webpage that match an extraction pattern are outlined in red. The "my patterns" button opens a window that displays a list of extraction patterns that the user has used in the past (Figure 4). These patterns may be ones created by the user or downloaded from the community pattern repository. The user can delete or download new patterns through this interface. The "community patterns" button displays a list of extraction patterns that all users have created. Users can filter the list and look for all patterns created by a specific contributor. The "questions" button opens a mail client and allows users to submit questions at any time. The "info" button takes the user to the study website. And, the "feedback" button opens a survey that allows the user to submit feedback at any time.

Community pattern repository

Every time a user creates an extraction pattern, it is added to the community pattern repository. This repository is in turn visible to all of the participants. Figure 3 shows the interface for viewing all patterns that are applicable to the current page. Each pattern in the list is displayed according to its domain, its author, and the number of webpage elements it can collect on the current page. The display also lists the total number of webpage elements the pattern can extract. Thus, the first pattern listed in Figure 3 is for the www.yelp.com website, was created by a user named *larry*, and can collect seven webpage elements from this webpage for the "Caffe Presse" restaurant on yelp.com. It can, however, extract a total of up to eight webpage elements, which means that there is one more item that it can collect that either does not exist on this webpage or is in a different location. It is likely that *larry* created the extraction pattern for yelp.com on a different webpage, not the one for "Caffe Presse," which is why it only matches partially. Patterns that only match partially may be more versatile and account for layout variations. Alternatively, webpage structure changes may make it difficult to extract all originally selected elements. The patterns are listed according to highest number of extractable webpage elements. To view the webpage elements that are extracted by a pattern, the user moves the cursor over the list of patterns, and the elements that can be extracted by the pattern underneath the cursor are highlighted in blue. In Figure 3 the cursor is over the first extraction pattern and seven webpage elements are highlighted, including the name, rating, restaurant categories, address, and reviews. To download a pattern the user clicks on the green check mark button ✓ and the pattern becomes part of his personal pattern repository.



Unverified Extraction Patterns

Community Patterns

www.yelp.com - user: larry - matches: 7/8
www.yelp.com - user: ajaffe - matches: 6/9
www.yelp.com - user: rodrig - matches: 5/6
www.yelp.com - user: nbn - matches: 5/6
www.yelp.com - user: mistermartin - matches: 5/5
www.yelp.com - user: Anonymous - matches: 4/5
www.yelp.com - user: alex - matches: 4/5
www.yelp.com - user: smc - matches: 4/8
www.yelp.com - user: mjo - matches: 4/4
www.yelp.com - user: brian - matches: 4/9

OK Cancel

Cafe Presse
based on 52 reviews Rating Details »
Categories: French, Breakfast [Edit]
1117 12th Ave
(between Madison St & Spring St)
Seattle, WA 98122
(206) 709-7674
www.cafepresseseattle.com

Hours: Tue 7:00 a.m.-2:00 a.m.	Price Range: $$
Good for Groups: Yes	Parking: Street
Delivers: No	Good for Kids: No
Wheelchair Accessible: Yes	Take-out: Yes
BYOB: No	Outdoor Seating: No
	Corkage: Yes

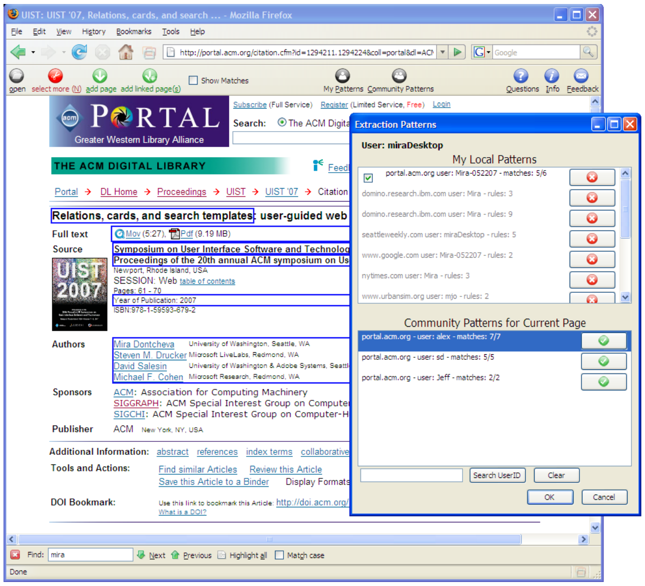
52 Reviews for Cafe Presse
Sort by: Recent + Votes | Time | Rating
All Reviews

Jinny K.
Reviews Written: 9
10/29/2007
I find it difficult in Seattle to find an out-of-the-ordinary breakfast menu and even more difficult to find it executed well--Cafe Presse delivers on both fronts. Pomme frites with their seasoned mayo were tasty. Even the hard-boiled eggs were done right (soft yolks). Plus, how can you go wrong with adding mayo to everything? No scrimping and saving on calories here...hooray! They serve Vita coffee, which is nice, but I prefer coffee at the adjacent Stumptown. And, if I've failed to regale and convince you with my sumptuous stories of mayo...even their baguettes were yummy. Go, immediately.

People Who Viewed This Also
Le Pichet
74 reviews
Neighborhood: Down
Category: French
DIDA
1 review
Neighborhood: Greer
Dinette
26 reviews
Neighborhood: Capit
Category: French
611 Supreme

Figure 3. When the "add page" button is purple it opens the interface to the community pattern repository. The user can view all patterns that apply to the current webpage and download them. To see which parts of the page are extracted by a pattern, the user moves the cursor over the list of patterns. The elements that can be extracted by the pattern under the cursor are highlighted in blue.

Users can view patterns they have used in the past by clicking on the "my patterns" button, which opens a view of their personal pattern repository. Figure 4 shows the interface for the personal pattern repository. It includes two lists. The top list shows the patterns that the subject has used previously. To delete a pattern the user can press on the delete button. All patterns that do not apply to the current webpage appear grey. The checkbox allows a user to store many


UIST: UIST '07, Relations, cards, and search ... - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://portal.acm.org/citation.cfm?id=1294211.1294224&coll=portal&dl=ACM
Google
open
select more (N)
add page
add linked page(s)
Show Matches
My Patterns
Community Patterns
Questions
Info
Feedback
Subscribe (Full Service) Register (Limited Service, Free) Login
Search: The ACM Digita
PORTAL
Greater Western Library Alliance
THE ACM DIGITAL LIBRARY
Feed
Portal → DL Home → Proceedings → UIST → UIST '07 → Citation
Find: mira
Next
Previous
Highlight all
Match case
Done

Relations, cards, and search templates: user-guided web

Full text Mov (5:27), Pdf (9.19 MB)

Source Symposium on User Interface Software and Technolog
Proceedings of the 20th annual ACM symposium on Us
Newport, Rhode Island, USA
SESSION: Web table of contents
Pages: 61 - 70
Year of Publication: 2007
ISBN:978-1-59593-679-2

Authors	
Mira Dontcheva	University of Washington, Seattle, WA
Steven M. Drucker	Microsoft LiveLabs, Redmond, WA
David Salesin	University of Washington & Adobe Systems, Seattl
Michael F. Cohen	Microsoft Research, Redmond, WA

Sponsors ACM: Association for Computing Machinery
SIGGRAPH: ACM Special Interest Group on Compute
SIGCHI: ACM Special Interest Group on Computer-H

Publisher ACM New York, NY, USA

Additional Information: abstract references index terms collaborative

Tools and Actions: Find similar Articles Review this Article
Save this Article to a Binder Display Formats

DOI Bookmark: Use this link to bookmark this Article: http://doi.acm.org/
What is a DOI?

Extraction Patterns

User: miraDesktop

My Local Patterns

☑ portal.acm.org user: Mira-052207 - matches: 5/6
domino.research.ibm.com user: Mira - rules: 3
domino.research.ibm.com user: Mira - rules: 9
seattleweekly.com user: miraDesktop - rules: 5
www.google.com user: Mira-052207 - rules: 2
nytimes.com user: Mira - rules: 3
www.urbansim.org user: mjo - rules: 2

Community Patterns for Current Page

portal.acm.org - user: alex - matches: 7/7 ✓
portal.acm.org - user: sd - matches: 5/5 ✓
portal.acm.org - user: Jeff - matches: 2/2 ✓

Figure 4. The user can view the patterns he has created and downloaded by pressing the "my patterns" button, which shows a list of personal patterns and a list of community patterns that apply to the current page.

patterns for a particular webpage and change which one is considered default. The second list shows patterns that are available for the current webpage in the community repository. As the user moves the cursor over the list of patterns, the webpage elements that can be extracted by the pattern under the cursor are highlighted. The user can at any time add a pattern to his personal pattern repository by clicking on the check mark button ✓.

STUDY METHODOLOGY

We recruited 24 participants through emails to university distribution lists. The subjects were offered a gift certificate to the university bookstore for their participation. The gift certificate was prorated based on their participation. Ten of the participants were female and fourteen were male. The participants ranged in age from 19 to 67, with a median age of 28 and a mean age of 31. Fourteen of the participants were graduate students in technical departments; four were undergraduates in technical departments; and six were staff members at the university. Nineteen of the participants stated that they perform some form of Web research every day, while the remaining participants said they perform such tasks several times a week. When questioned about which tools they typically use, twenty of the participants mentioned emailing themselves and using bookmarks. Sixteen said that they copy and paste information into a document. Thirteen said they print webpages. Ten said they send email to other people. Nine mentioned more advanced tools such as specialized toolbars or online tools such as Google Notebook and `del.icio.us`. When asked about how often they organize their Web content collections, half of the participants said they either never organize Web content or only do so as needed and when time allows. A third of the participants said they organize things when they collect new content. Only two participants mentioned organizing content several times a month, and one stated organizing content several times a week. Thus, our participant population showed behavior similar to user behaviors described in previous studies [2] [11] – frequent Web use for research with infrequent organization.

The study included two in-person interviews, one at the beginning of the study and one at the end. During the first interview, the participants filled out a demographic questionnaire and took part in a tutorial. We installed the Web Summaries extension on their computer, showed them how to use the tool, and then gave them a specific task to test whether they understood the tool. All of the participants used either a personal laptop or a work desktop computer for the study. The subjects were instructed to use the Web Summaries tool for their own personal tasks. During the closing interview, the participants filled out a closing survey while we uninstalled Web Summaries. In addition to the two interviews, the study included weekly surveys, which asked about their experiences with the tool. The participants were also assigned weekly tasks, which asked them to visit a specific website and collect information from it. In total the users received seven tasks, including:

- **Shopping** – visit `amazon.com` and collect 10 books of your choice.
- **Reviews** – visit `yelp.com` and collect 10 items of your choice.
- **Travel** – visit `tripadvisor.com` and collect information about 10 different hotels.
- **Events** - visit `upcoming.org` and collect 10 different upcoming events for the city of your choice.
- **Entertainment** – visit `imdb.com` and collect information about 10 movies and/or actors.
- **Cooking** – visit `epicurious.com` and collect 10 recipes.
- **Reference** – visit the ACM Digital Library at `portal.acm.org` and collect references to 10 articles.

The goal in assigning tasks was to give the participants examples of tasks for which they might consider using Web Summaries. Additionally, the weekly tasks were useful in simulating a large number of participants by encouraging all subjects to visit the same website with the same goal in mind, thereby allowing us to evaluate the community pattern repository interface. In addition to soliciting user feedback through surveys, Web Summaries also logged all user interaction with the extension such as toolbar button clicks and interactions with the summary items and views. The log files were uploaded directly to the server without any user intervention.

RESULTS

Of the 24 initial participants, only 15 were active contributors and completed the study. Thirteen participants took part in the closing interviews. An additional two participants were frequent users of the tool but did not participate in the closing interviews. We received 60 survey responses

event	number logged	type
loaded new webpage	15009	browsing
switched tab	1854	
created new pattern	257	toolbar
modified pattern	138	
pressed add page button	425	
pressed add linked pages button	254	
pressed my patterns button	95	
pressed community patterns button	66	
downloaded community pattern	54	
pressed information button	21	
pressed feedback button	9	
pressed questions button	8	
added new tag	298	tags and categories
added new category of tags	48	
deleted tag	29	
deleted category of tags	15	
added new item	1033	summary
changed view	318	
deleted item	235	
filtered using a facet	96	
clicked on dynamic update button	77	
clicked on link in summary	72	
assigned user tag	14	
removed filter	9	
created summary tag	3	
deleted summary tag	1	

Table 1. This table shows the frequency and types of logged events.

– 36 over email, and 24 through the tool. Figure 5 shows a sorted list of the participants according to the number of days each one used the Web Summaries tool. This data does not impose a minimum usage time, thus even if the user only opened Web Summaries briefly, we record that day as a day of usage. Figure 6 shows a temporal plot of the tool usage. The stacked bar plot shows the duration in minutes of each user's interaction with the tool for a given day. Each subject is represented by a different color. We compute the usage duration by ignoring browsing events and removing gaps in usage greater than 20 minutes. The grey vertical lines are the dates on which the participants received a new task. The graph shows active initial usage that decreased over time. We expect that this is due to the initial novelty of the tool and an initial period of exploration. While some participants only used the tool for the assigned tasks, many participants found personal tasks for which it was useful.

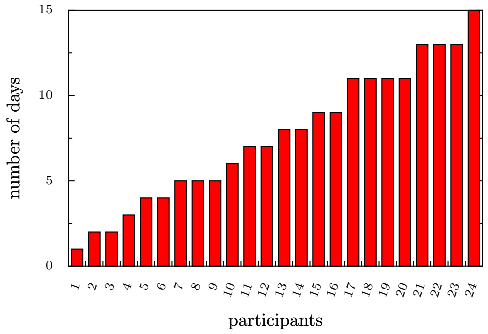


Figure 5. This figure shows the number of days each participant used the Web Summaries tool.

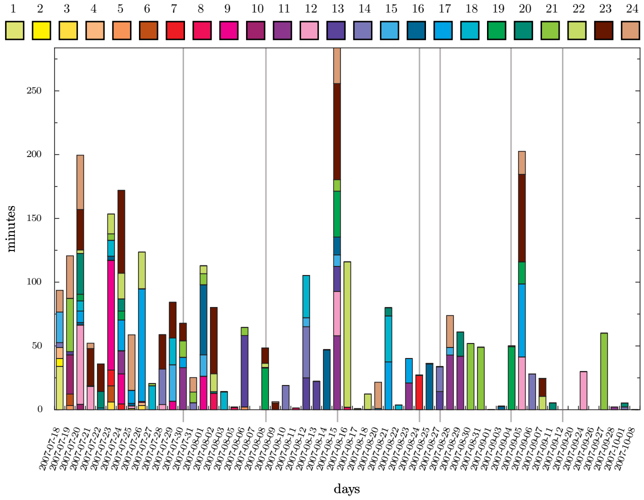


Figure 6. This figure shows the usage of the Web Summaries tool over time. The grey lines correspond to dates on which the participants received a task. Each color corresponds to a study participant.

During the ten-week study, Web Summaries logged 26244 events. The tool only logged user activities when it was active (i.e., when the Web Summaries window was open). It logged user interactions with the summary, pattern specification and editing, and all toolbar button clicks. The tool also logged the URL of every webpage the user visited. Table 1 shows the frequency and types of events that were logged. The events are grouped into four types: browsing, toolbar, tags and categories, and summary. *Browsing events* include switching tabs and loading a new webpage. The majority of the logged events were of this type. *Toolbar events* include pattern specification events and any toolbar button presses, such as pressing the "add page" button or "add linked pages" button. The *tags and categories events* include creating and deleting tags and categories. In the interface categories were listed as groups. *Summary events* include all user actions on the actual summary such as changing views or deleting an item. Some events were logged for debugging purposes. For example, the tool logged 4134 pattern specification events that show all user clicks on a page during pattern specification.

The data shows that the participants created many patterns and used them repeatedly to collect over 1000 items of content from different websites. They also created their own tags for the content that were personally meaningful to them. The participants changed summary views often and deleted content they did not find useful. However, they did very little organization of the summaries. The majority of the log events were browsing events because many times users left the summary window open and continued browsing for some other purpose.

Qualitative feedback

During the closing interviews we asked the participants about their favorite features of Web Summaries. Half of the participants said that the automatic gathering of content through the "add linked pages" button was their favorite feature. Other favorite aspects included the ability to save persistent copies of Web information that is often dynamic, the ability to collect text and images and make rich views of the content, the variety of views offered by Web Summaries, the tool's integration with the browser, and the fact that only one tool was necessary for accomplishing a variety of tasks. Users wrote:

> *"I liked constructing 'captures' - the interaction was easy and fun and I liked that I could customize what I wanted to capture. I liked the way the 'captures' are displayed in the thumbnail view."*

> *"I liked its intuitive interface and close integration with the existing page."*

> *"I love my 'add linked' pages option! It's such a time-saving functionality. It makes the process so easy once you've decided what info you want to gather."*

> *"Add linked pages tool = HUGE time saver."*

When we tried to uninstall the tool, two of the participants asked to continue to use it so that they could access the data they had collected. One of the participants used the tool for her own work and said that it helped her conduct her own research. She said,

> *"I found that Web Summaries are great at saving my time of looking up, revisiting, and documenting Web-SVN and bug databases. The Web Summary that I created is now a part of my research document."*

Although users were very positive about the automatic gathering functionalities, they also had many suggestions for improvement. The participants requested more feedback during the automatic extraction process, a more flexible interface for specifying and modifying extraction patterns, a smaller toolbar, and the ability to edit the summary views. During the closing interviews some of the participants remarked that they had used the tool much less than they originally expected. Many users had small displays and the size of the summary toolbar limited their browser screen real estate. As a result they hid the Web Summaries toolbar and often forgot about the availability of the tool. Since the users had to look at the summary window to see if the content they wanted to store had been added to their summary, many reported spending too much time switching between the browser window and the summary window. We expect that users with multi-monitor displays will have a much easier time with the summary window, but in order to support laptop users we plan to explore different approaches for providing feedback. Despite an imperfect interface, many of the users completed the assigned tasks and did use the tool for personal tasks.

Collecting Web content with extraction patterns

During the ten-week study, the participants created 257 extraction patterns, used the "add page" button 425 times and collected information from 987 hyperlinks by applying the "add linked pages" functionality 254 times. They created a total of 447 unique summaries. Figure 7 shows the number of patterns created over time. The grey lines signify the dates on which the participants received a new task. The participants also modified extraction patterns 138 times. Of the 257 patterns, 114 were used to collect content automatically. The remaining 147 extraction patterns were not applied to automatically collect content, because their authors were interested in clipping content rather than collecting a number of comparable items. Patterns were used on average 9 times with a standard deviation of 10. Figure 8 shows the utility of each pattern. Most of the patterns that were used more than 10 times were shared among the participants. Interestingly, the pattern that was used the most, over 80 times, was used by only one participant. It was created for the ACM Digital Library website and used to collect over 80 references.

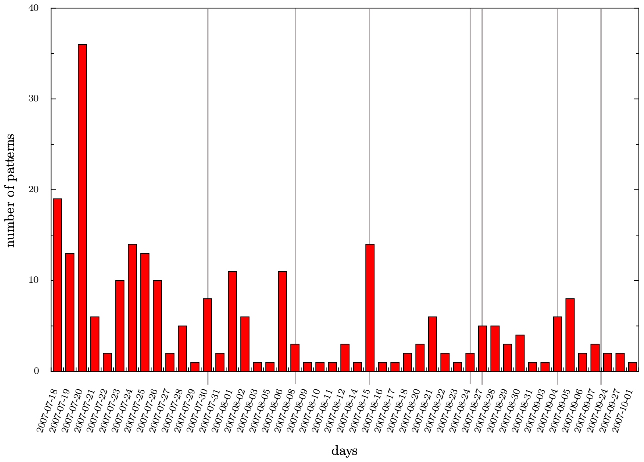


Figure 7. This figure shows the number of extraction patterns created during the study. The grey lines correspond to dates on which the participants received a task.

Despite the fact that the subjects created and used many patterns, they encountered some problems when creating patterns. The modal interface for specifying extraction patterns was sometimes confusing. Several users requested that the tool automatically save their selection when they navigate away from the page or close the tab. Several users requested a more flexible selection mechanism.

> *"It works pretty well, but there are times when I would rather have more control over what's selected, i.e, combining elements into one or selecting just the part of an element that's after certain punctuation."*

One participant who is a Web programmer requested an interface for specifying regular expressions as extraction patterns. This would have enabled him to collect content from the website `mybus.org`. Structural extraction patterns are not appropriate for this website as the location of the buses changes according to arrival time. Another participant requested retrieval from a list of items, as many website do not have detail webpages for particular events or locations. Other participants requested multi-page extraction patterns, as content is often split among several tabs or screens.

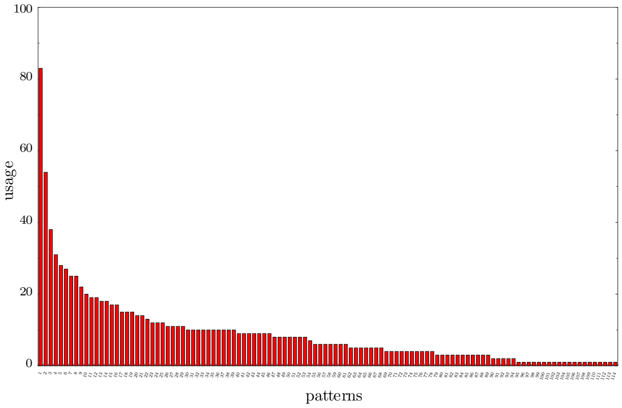


Figure 8. This figure shows the utility of each pattern.

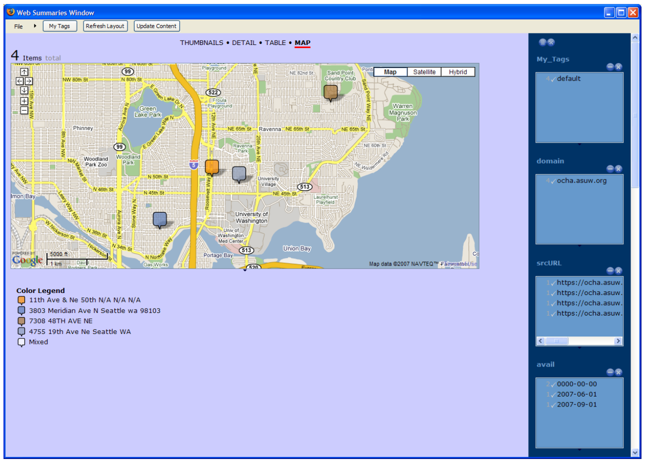
type	website	participants	visits
shopping	`amazon.com`	12	14
reviews	`yelp.com`	14	14
travel	`tripadvisor.com`	14	16
events	`upcoming.org`	11	11
entertainment	`imdb.com`	12	14
cooking	`epicurious.com`	8	10
reference	`portal.acm.org`	4	5

Table 2. This table shows subject participation in the assigned tasks. The visits column shows that some participants visited the website more than once.

Accomplishing tasks

Most of the active study subjects completed the assigned tasks. Table 2 shows the number of subjects that visited the assigned task websites. Some of the users visited the assigned task websites more than once. The last column shows the discrete visits to those websites. The low visit rate for the reference task is due to the fact that some participants did not use the ACM Digital Library for academic references. They chose to use other reference libraries such as Inspec or IEEE Xplore. Also, since this was the last assigned task, the participants did not have as much time to complete the task.

In addition to successfully accomplishing many of the assigned tasks, the subjects also used Web Summaries for a number of personal tasks. The participants created extraction patterns on 88 distinct domains and used automatic extraction on 44 of those domains. They collected a variety of data for many purposes. We group task data into three categories - life, work, and fun. Most of the participants engaged in some type of life information task, such as comparison shopping (Figure 1), searching through rentals (Figure 9) or job listings (Figure 10), looking for local car washes or farms. Work information tasks included collecting information about conference courses, saving articles (Figure 12), and gathering technical data (Figure 11). The participants also used Web Summaries to collect information for hobbies. One participant collected comics (Figure 13) and used the "update content" button to automatically retrieve the most recent comic strip. Another participant collected information about local hiking trails, while yet another collected historical weather data (Figure 14). These sample summaries show that users collected a large variety of data. Some collected highly structured content, while others stored entire articles with many paragraphs of text.



4 Items total

Color Legend
- 11th Ave & Ne 50th N/A N/A N/A
- 3803 Meridian Ave N Seattle wa 98103
- 7308 48TH AVE NE
- 4755 19th Ave Ne Seattle WA
- Mixed

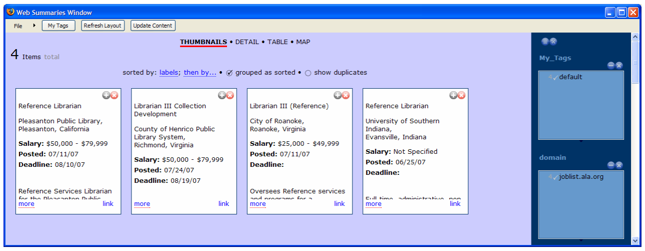
My_Tags: 4 ✓ default

domain: 4 ✓ ocha.asuw.org

srcURL:
1 ✓ https://ocha.asuw.
1 ✓ https://ocha.asuw.
1 ✓ https://ocha.asuw.
1 ✓ https://ocha.asuw.

avail:
2 ✓ 0000-00-00
1 ✓ 2007-06-01
1 ✓ 2007-09-01

Figure 9. One participant used Web Summaries to collect advertisements for vacant apartments.



4 Items total

sorted by: labels; then by... • ☑ grouped as sorted • ○ show duplicates

Reference Librarian

Pleasanton Public Library,
Pleasanton, California

Salary: $50,000 - $79,999
Posted: 07/11/07
Deadline: 08/10/07

Reference Services Librarian
for the Pleasanton Public

Librarian III Collection Development

County of Henrico Public
Library System,
Richmond, Virginia

Salary: $50,000 - $79,999
Posted: 07/24/07
Deadline: 08/19/07

Librarian III (Reference)

City of Roanoke,
Roanoke, Virginia

Salary: $25,000 - $49,999
Posted: 07/11/07
Deadline:

Oversees Reference services
and programs for a

Reference Librarian

University of Southern
Indiana,
Evansville, Indiana

Salary: Not Specified
Posted: 06/25/07
Deadline:

Full time, administrative, non

My_Tags: 4 ✓ default

domain: 4 ✓ joblist.ala.org

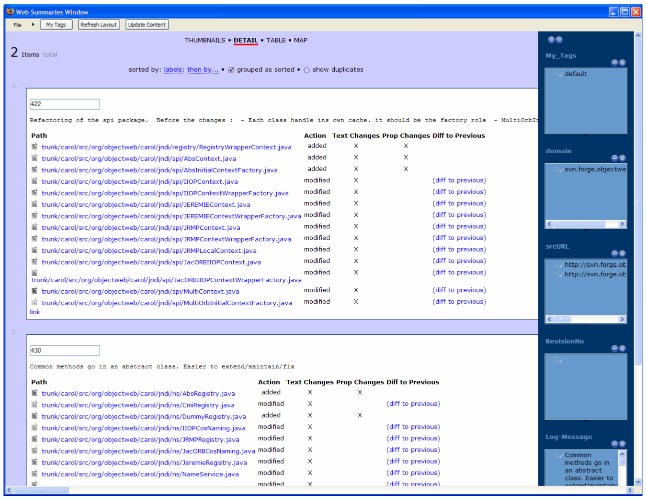
Figure 10. Another participant collected listings for job openings.

Organizing summaries

In addition to automatically gathering content, users had the opportunity to explore a richer organization metaphor than is currently possible in the browser through layout templates. During the ten-week study, the participants created 447 unique summary collections and added over 1000 items to those collections. Figure 15 shows the number of summaries each participant created. Most of the summaries were not accessed more than once, which is in line with the participants' survey feedback. Although the participants used Web Summaries for a variety of tasks, they did not return to their summaries and continue with the tasks at a later time. It is likely that the subjects did not encounter a content intensive task that required returning to a summary during the study period. Also, since the participants knew that this was a study, they may have been hesitant to store too much important data in a format that may not be accessible later.

Many users mentioned the ability to store rich information from webpages.

> *"It let me collect richer info about the sites I visited than the text files I usually use to take notes. I liked having images and active links, for example."*



2 Items total

sorted by: labels; then by... • ☑ grouped as sorted • ○ show duplicates

1.

422

Refactoring of the spi package. Before the changes : - Each class handle its own cache. it should be the factory role - MultiOrbI

Path	Action	Text Changes	Prop Changes	Diff to Previous
trunk/carol/src/org/objectweb/carol/jndi/registry/RegistryWrapperContext.java	added	X	X	
trunk/carol/src/org/objectweb/carol/jndi/spi/AbsContext.java	added	X	X	
trunk/carol/src/org/objectweb/carol/jndi/spi/AbsInitialContextFactory.java	added	X	X	
trunk/carol/src/org/objectweb/carol/jndi/spi/IIOPContext.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/IIOPContextWrapperFactory.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JEREMIEContext.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JEREMIEContextWrapperFactory.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JRMPContext.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JRMPContextWrapperFactory.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JRMPLocalContext.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JacORBIIOPContext.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/JacORBIIOPContextWrapperFactory.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/MultiContext.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/spi/MultiOrbInitialContextFactory.java	modified	X		(diff to previous)

link

2.

430

Common methods go in an abstract class. Easier to extend/maintain/fix

Path	Action	Text Changes	Prop Changes	Diff to Previous
trunk/carol/src/org/objectweb/carol/jndi/ns/AbsRegistry.java	added	X	X	
trunk/carol/src/org/objectweb/carol/jndi/ns/CmiRegistry.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/ns/DummyRegistry.java	added	X	X	
trunk/carol/src/org/objectweb/carol/jndi/ns/IIOPCosNaming.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/ns/JRMPRegistry.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/ns/JacORBCosNaming.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/ns/JeremieRegistry.java	modified	X		(diff to previous)
trunk/carol/src/org/objectweb/carol/jndi/ns/NameService.java	modified	X		(diff to previous)
	modified	X		(diff to previous)

My_Tags
2 ✓ default

domain
2 ✓ svn.forge.objectwe

srcURL
1 ✓ http://svn.forge.ob
1 ✓ http://svn.forge.ob

RevisionNo
2 ✓

Log-Message
1 ✓ Common methods go in an abstract class. Easier to extend/maintain/

Figure 11. One participant used the Web Summaries tool in her own research and collected information about open source CVS repositories.

> *"I liked the fact that in addition to text, I could tag pictures, and this made the thumbnail view really useful: I could see what recipes were for rather than just read the title and that made browsing easier."*

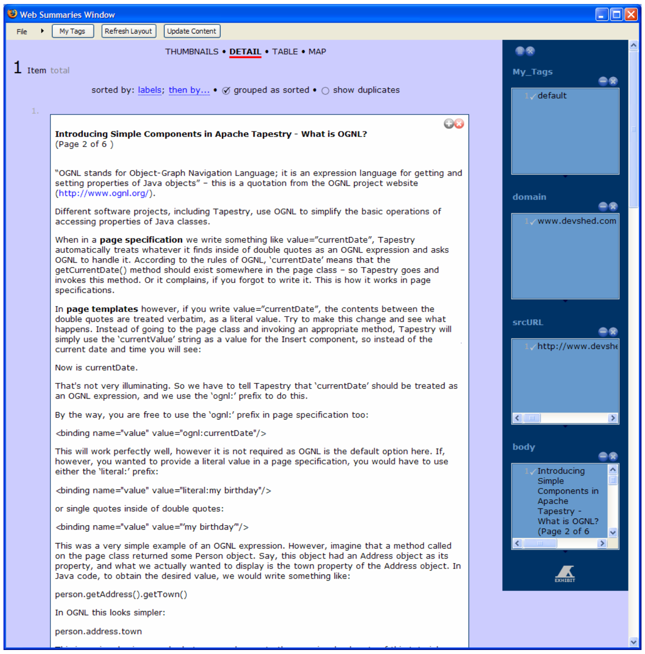
While the log data shows that the participants were not accessing existing summaries often, many of the participants asked to be informed of future releases. Two of the participants requested that they continue using the tool so that they could continue with the summaries they had created.

The log data shows that the participants often changed views. Web Summaries recorded 318 view change events. Of those 318 events, 104 were changes to the detail view; 98 were changes to the table view; 70 were changes to the thumbnail view; and 46 were changes to the map view. Since the thumbnail view was the default view, it is largely underrepresented in these statistics. User feedback on favorite views varied according to the person and task.

> *"At first I preferred the table view, because it presented all the data side by side for comparison. Then, as the summaries became more complex, and there was no way (that I could find) to resize columns, change font size, etc., I started using the thumbnail view instead."*

> *"Table works well when there are lots of items. Thumbnail works well for a few items when the first few fields contain crucial information."*

Table 1 shows that users deleted items in the summary frequently. This number is likely inflated because of the duplication of content during pattern modification. When users modify an extraction pattern, the system adds another item to their collection with the newly created pattern. The participants filtered their summaries often but they did very little organization through tagging. It is possible that since the participants had already tagged the content during the clipping phase, they didn't feel they need to tag the content collections. This minimal amount of organization could also be due to the small sizes of the summary collections. The summaries included on average ten items. Finally, users used the dynamic update functionality and clicked on links stored in the summary frequently.



1 Item total

sorted by: labels; then by... • ☑ grouped as sorted • ○ show duplicates

1.

Introducing Simple Components in Apache Tapestry - What is OGNL?
(Page 2 of 6)

"OGNL stands for Object-Graph Navigation Language; it is an expression language for getting and setting properties of Java objects" – this is a quotation from the OGNL project website (http://www.ognl.org/).

Different software projects, including Tapestry, use OGNL to simplify the basic operations of accessing properties of Java classes.

When in a **page specification** we write something like value="currentDate", Tapestry automatically treats whatever it finds inside of double quotes as an OGNL expression and asks OGNL to handle it. According to the rules of OGNL, 'currentDate' means that the getCurrentDate() method should exist somewhere in the page class – so Tapestry goes and invokes this method. Or it complains, if you forgot to write it. This is how it works in page specifications.

In **page templates** however, if you write value="currentDate", the contents between the double quotes are treated verbatim, as a literal value. Try to make this change and see what happens. Instead of going to the page class and invoking an appropriate method, Tapestry will simply use the 'currentValue' string as a value for the Insert component, so instead of the current date and time you will see:

Now is currentDate.

That's not very illuminating. So we have to tell Tapestry that 'currentDate' should be treated as an OGNL expression, and we use the 'ognl:' prefix to do this.

By the way, you are free to use the 'ognl:' prefix in page specification too:

<binding name="value" value="ognl:currentDate"/>

This will work perfectly well, however it is not required as OGNL is the default option here. If, however, you wanted to provide a literal value in a page specification, you would have to use either the 'literal:' prefix:

<binding name="value" value="literal:my birthday"/>

or single quotes inside of double quotes:

<binding name="value" value="'my birthday'"/>

This was a very simple example of an OGNL expression. However, imagine that a method called on the page class returned some Person object. Say, this object had an Address object as its property, and what we actually wanted to display is the town property of the Address object. In Java code, to obtain the desired value, we would write something like:

person.getAddress().getTown()

In OGNL this looks simpler:

person.address.town

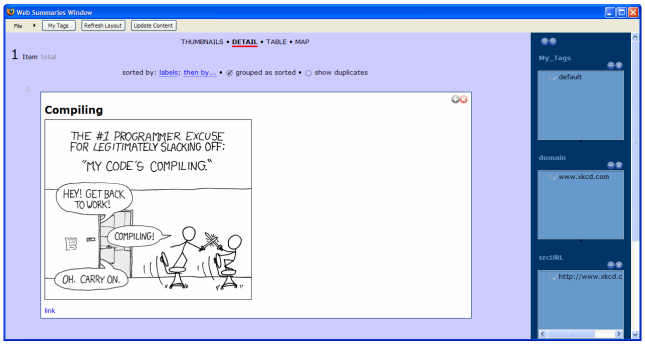
My_Tags
1 ✓ default

domain
1 ✓ www.devshed.com

srcURL
1 ✓ http://www.devshe

body
1 ✓ Introducing Simple Components in Apache Tapestry - What is OGNL? (Page 2 of 6

Figure 12. Some participants collected entire articles from the Web.



1 Item total

sorted by: labels; then by... • ☑ grouped as sorted • ○ show duplicates

1.

Compiling

link

My_Tags
1 ✓ default

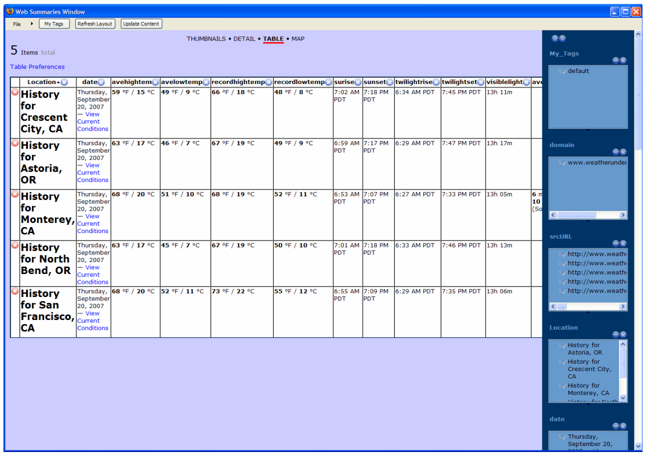
domain
1 ✓ www.xkcd.com

srcURL
1 ✓ http://www.xkcd.c

Figure 13. One participant collected comics and used automation to automatically collect the most recent comic strip.

Sharing patterns with the community pattern repository

One of the goals of this study was to better understand the challenges surrounding a public repository of extraction patterns. Our analysis reveals that users are interested in using a collaborative pattern repository but that the interface for conveying which patterns are available and what type of information they can extract needs to be finely tuned to make it easy to select an appropriate pattern. In total there were 54 downloads of 30 unique patterns from the community repository with only 9 downloads of patterns by their original authors. Ten patterns were downloaded more than once and all of the patterns that were reused were for the assigned task websites. There was very little overlap in browsing habits between the 15 active participants thus the assigned tasks served us well in simulating a larger participant population.



	Location	date	avehightem	avelowtemp	recordhightemp	recordlowtemp	surise	sunset	twilightrise	twilightset	visiblelight	ave
	History for Crescent City, CA	Thursday, September 20, 2007 — View Current Conditions	59 °F / 15 °C	49 °F / 9 °C	66 °F / 18 °C	48 °F / 8 °C	7:02 AM PDT	7:18 PM PDT	6:34 AM PDT	7:45 PM PDT	13h 11m	
	History for Astoria, OR	Thursday, September 20, 2007 — View Current Conditions	63 °F / 17 °C	46 °F / 7 °C	67 °F / 19 °C	49 °F / 9 °C	6:59 AM PDT	7:17 PM PDT	6:29 AM PDT	7:47 PM PDT	13h 17m	
	History for Monterey, CA	Thursday, September 20, 2007 — View Current Conditions	68 °F / 20 °C	51 °F / 10 °C	68 °F / 19 °C	52 °F / 11 °C	6:53 AM PDT	7:07 PM PDT	6:27 AM PDT	7:33 PM PDT	13h 05m	6 m 10 (So
	History for North Bend, OR	Thursday, September 20, 2007 — View Current Conditions	63 °F / 17 °C	45 °F / 7 °C	67 °F / 19 °C	50 °F / 10 °C	7:01 AM PDT	7:18 PM PDT	6:33 AM PDT	7:46 PM PDT	13h 13m	
	History for San Francisco, CA	Thursday, September 20, 2007 — View Current Conditions	68 °F / 20 °C	52 °F / 11 °C	73 °F / 22 °C	55 °F / 12 °C	6:55 AM PDT	7:09 PM PDT	6:29 AM PDT	7:35 PM PDT	13h 06m	

Figure 14. One participant collected detailed historical weather data.

Many participants downloaded a pattern and then modified it to fit their preferences. Users said:

> *“I loved when I could use other people’s extraction pattern. These were a bit mysterious regarding when they would appear. Often I used other peoples’ as a base and revised them to my liking.”*

> *“I liked the way I could hover over a pattern in the list and see which elements on the page were selected.”*

The types of patterns users would create and use was somewhat personal. Some preferred collecting lots of information, others preferred collecting less information.

> *”I tended to look for patterns with the most matches. This may not have left me with the ”best” pattern for the page, though.”*

> *“I checked first to see if there are some good pattern out there. If not I usually create one myself. I want a simple one, not with a lot of information.”*

Some users preferred to create their own patterns because they found the community pattern repository interface confusing. Some simply found it faster to create their own patterns.

> *“It seems like work to figure out what patterns are there and whether I want to use them. And making patterns still feels fun and easy so I’d rather make my own.”*

> *“When I tried using existing patterns I ran into some difficulties and it turned out to be faster to construct my own.”*

Despite a small and fairly private community, some participants didn’t trust the community patterns and preferred to create their own.

> *“I do my own thing because I don’t trust other people. Sharing is nice if you are in a hurry, but I usually created my own for personal reasons.”*

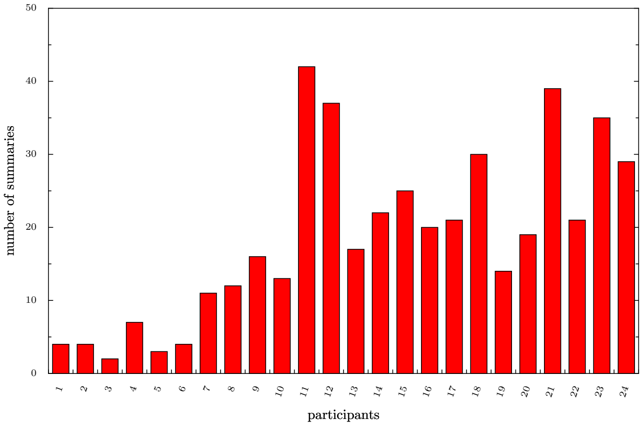


Figure 15. This figures shows the number of summaries created by each participant.

> *“I generally create my own. Creating extraction patterns is not that hard - it’s actually much harder to figure out what other people’s extraction patterns are. If I trusted other people’s extraction patterns, I would probably use them more.”*

Many participants found it easier to create their own patterns rather than learn how to use the community pattern repository interface. Participants who were in a hurry didn’t have much patience for the interface and preferred to interact with the already familiar clipping interface. When participants did find appropriate community patterns quickly, they used them.

DISCUSSION

Our field study shows that users enjoy creating extraction patterns and find automatic Web content extraction capabilities for structured Web content very useful. On many occasions users also collect less structured information, such as articles, and for such tasks automatic extraction is not necessary. Future development of tools and applications for automatically aggregating and extracting Web content needs to address the inherent user need for both structured and unstructured information. Good user interfaces for managing semantic Web content must provide more than just tables, grids, and maps. In future work, we plan to explore interaction techniques and visualization paradigms for a heterogeneous set of Web content that includes both highly structured information, such as the price and address of a hotel, and highly unstructured information, such as long personal reviews or descriptions of amenities.

For many of our subjects the timing of the study was highly critical. Users with content intensive tasks were much more willing to spend time with the tool and learn its interface. The participant who integrated Web Summaries into her own research was compelled to do so because of an impending conference deadline. Most daily tasks, however, are transient and short-lived, and the subjects did not find Web Summaries useful and easy to integrate into their everyday tasks. We believe this was due to the conscious upfront effort required by Web Summaries. The user must open the summary window and actively save content. We designed Web Summaries for content intensive tasks, thus it is not that surprising that it is not as well suited to transient tasks. However, users who do not use Web Summaries often, may turn it off and forget that the tool is available. Future studies on exploratory Web research tasks should be aware of the timing sensitivity that is inherent in this type of research, and Web content tools should be targeted to address the different types of tasks users experience.

Based on our observations we categorized tasks into four types - short transient, long transient, short permanent, and long permanent tasks. *Short transient* tasks are typically finished quickly, such as finding a birthday present for next week, or finding a restaurant for a night out. *Long transient* tasks include making more than one arrangement and possibly coordinating with others such as planning a vacation or work trip. They may also include learning about a new topic, such a gardening or a new health concern. *Short permanent* tasks are often also called monitoring Web tasks [12] and include reading news or blogs every day or checking favorite sport websites. Finally, *long permanent* tasks are tasks that are longstanding interests and involve gathering, collecting and organizing information over a long period of time. This categorization should viewed as a continuum. Some tasks start out as short transient tasks but may become longer transient tasks. Similarly, short transient tasks may become short permanent tasks. We designed Web Summaries for long transient and permanent tasks. We hope to adapt Web Summaries to shorter more transient tasks by allowing the user to retroactively build summaries thereby remove the active upfront need for managing content.

Finally, in our field study we explored the role of an online repository of extraction patterns. Such a repository could grow to become a collaborative user-defined Semantic Web. We found that when participants were in a hurry and wanted to quickly accomplish their task, they were more willing to use others' patterns. When their were not in a hurry or they had an important task, they created their own patterns. In the future we plan to explore visualization techniques and interfaces for exposing community information about a webpage to the user. An alternative to asking the user to select among many possible extraction patterns is to automatically select a good pattern using user preferences or statistics.

ACKNOWLEDGEMENTS

We thank our study participants for letting us analyze their information gathering activities. Cheyne Mathey-Owens was responsible for the server-side implementation of the Web Summaries system, and we appreciate his efforts. Many thanks to Susan Dumais for helping us define and focus this research study. Funding and research facilities were provided by Microsoft Research, Microsoft LiveLabs, Adobe Systems, and the University of Washington GRAIL lab.

REFERENCES

1. M. Bolin, M. Webber, P. Rha, T. Wilson, and R. C. Miller. Automation and customization of rendered web pages. In *Proc. of UIST*, pages 163–172, 2005.
2. H. Bruce, W. Jones, and S. Dumais. Keeping and re-finding information on the web: What do people do and what do they need to do? In *Proc. of ASIST*, 2004.
3. M. C. Daconta, L. J. Obrst, and K. T. Smith. *The Semantic Web: A Guide ot the Future of XML, Web Services, and Knowledge Management*. Wiley Publishing, Inc., 2003.
4. M. Dontcheva, S. M. Drucker, D. Salesin, and M. F. Cohen. Relations, cards, and search templates: user-guided web data integration and layout. In *Proc. of UIST*, pages 61–70, 2007.
5. M. Dontcheva, S. M. Drucker, G. Wade, D. Salesin, and M. F. Cohen. Summarizing personal web browsing sessions. In *Proc. of UIST*, pages 115–124, 2006.
6. J. Fujima, A. Lunzer, K. Hornbæk, and Y. Tanaka. Clip, connect, clone: combining application elements to build custom interfaces for information access. In *Proc. of UIST*, pages 175–184, 2004.
7. B. Hartmann, L. Wu, K. Collins, and S. R. Klemmer. Programming by a sample: rapidly creating web applications with d.mix. In *Proc. of UIST*, pages 241–250, 2007.
8. D. Huynh, S. Mazzocchi, and D. Karger. Piggy bank: Experience the semantic web inside your web browser. In *Proc. of ISWC*, 2005.
9. D. Huynh, R. Miller, and D. Karger. Exhibit: Lightweight structured data publishing. In *Proc. of WWW*, 2007.
10. D. F. Huynh, R. C. Miller, and D. R. Karger. Enabling web browsers to augment web sites' filtering and sorting functionalities. In *Proc. of UIST*, pages 125–134, 2006.
11. W. Jones, H. Bruce, and S. Dumais. Once found, what then?: A study of "keeping" behaviors in the personal use of web information. In *Proc. of ASIST*, 2002.
12. M. Kellar, C. Watters, and K. M. Inkpen. An exploration of web-based monitoring: implications for design. In *Proc. of SIGCHI*, pages 377–386, 2007.
13. E. Schrier, M. Dontcheva, C. Jacobs, G. Wade, and D. Salesin. Adaptive layout of dynamically aggregated documens. In *Proc. of IUI*, page (to appear), 2008.
14. A. J. Sellen, R. Murphy, and K. L. Shaw. How knowledge workers use the web. In *Proc. of the SIGCHI*, pages 227–234, 2002.
15. A. Sugiura and Y. Koseki. Internet scrapbook: automating web browsing tasks by demonstration. In *Proc. of UIST*, pages 9–18, 1998.
16. J. Wong and J. I. Hong. Making mashups with marmite: towards end-user programming for the web. In *Proc. of SIGCHI*, pages 1435–1444, 2007.