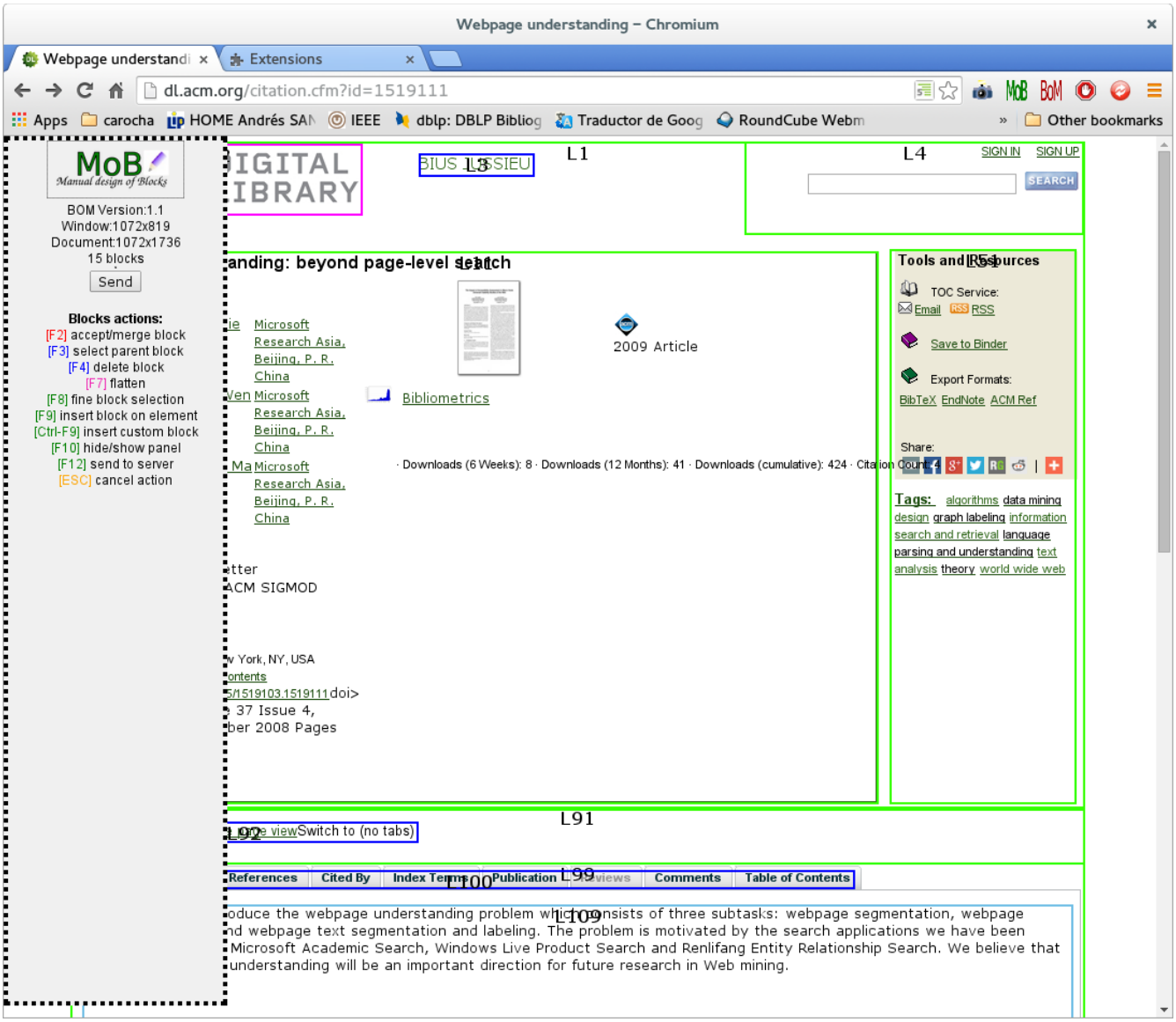
Manual Design of Blocks (MOB)
Quick Guide
Andrés Sanoja

The tool is operated by mouse clicks and keyboard function keys.

General preview of MOB in example web page (http://dl.acm.org/citation.cfm?id=1519111) with the granularity parameter value 2

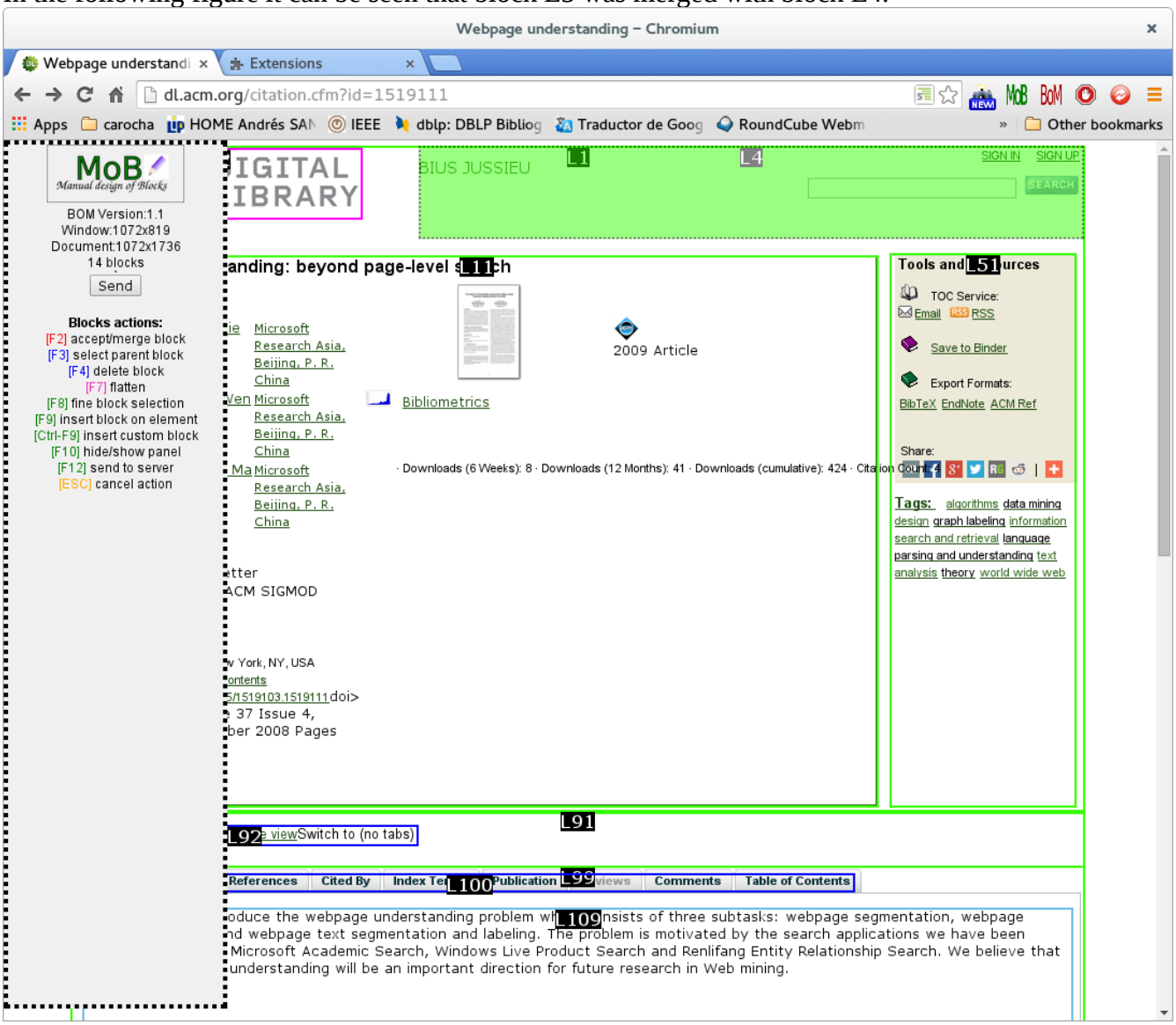


Accepting/Merging blocks

Accepting a block recompute all its properties and delete all children blocks. Selecting with the mouse two blocks (L3 and L4) and with the F2 key will merge both rectangles and accept the resulting block.

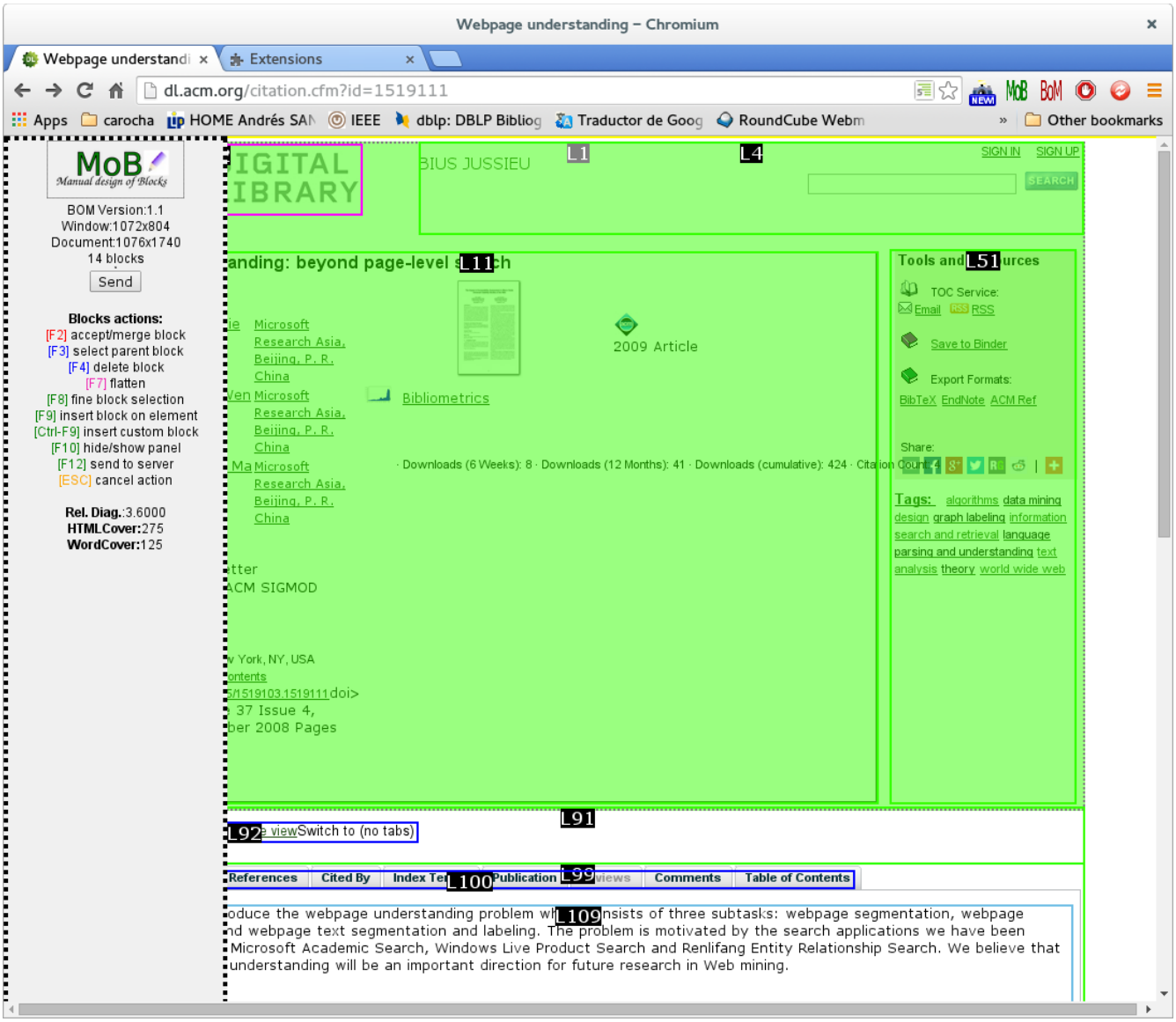
The result is one block which area is equivalent to that occupied by the two blocks.

In the following figure it can be seen that block L3 was merged with block L4.



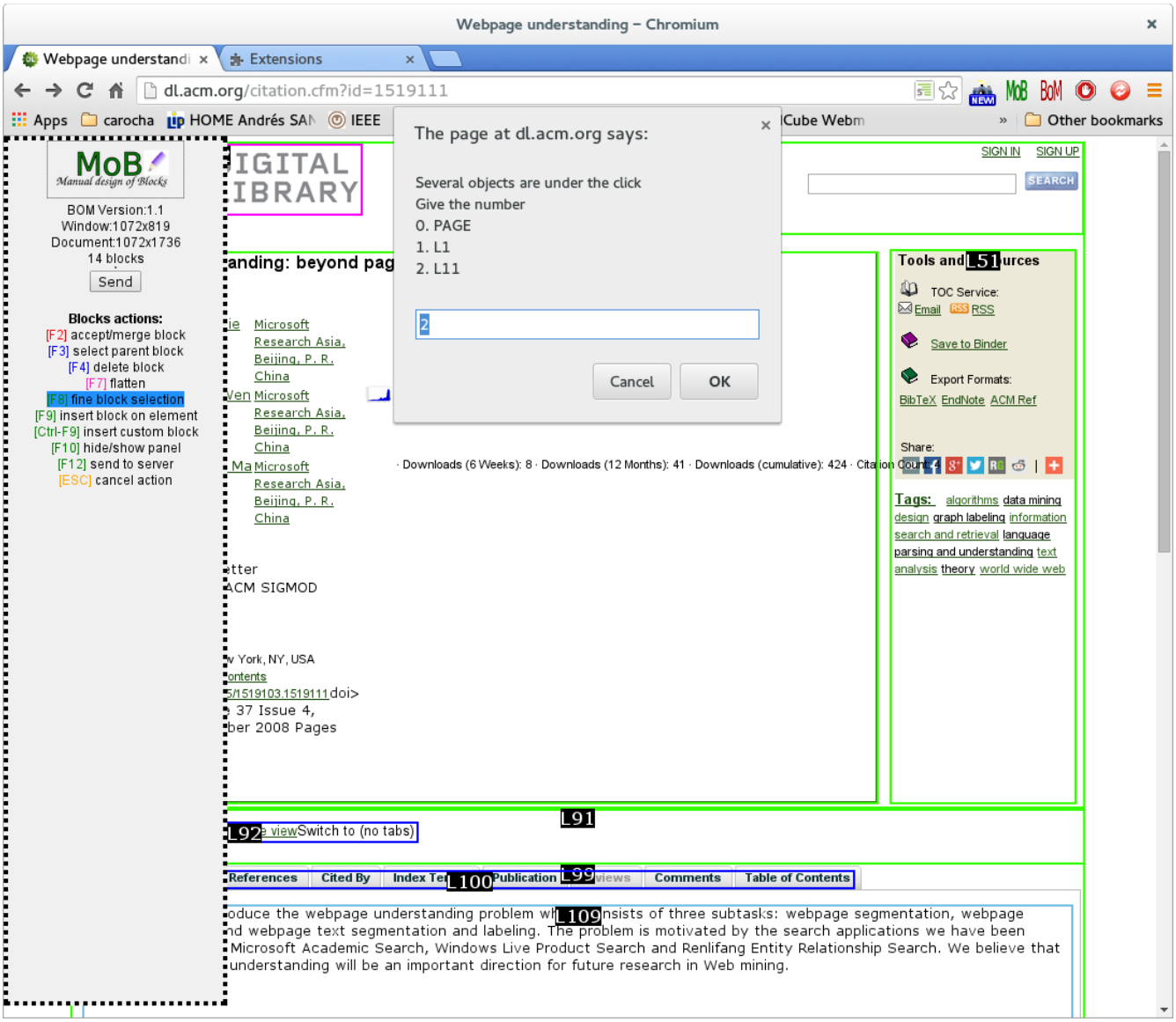
Navigating hierarchy

Selecting with the mouse one block (L4 for instance), we can go up to its parent by pressing F3. Select block L4 and pressing F3 with produce that the block L1 became the current block, as can be seen in the following figure.



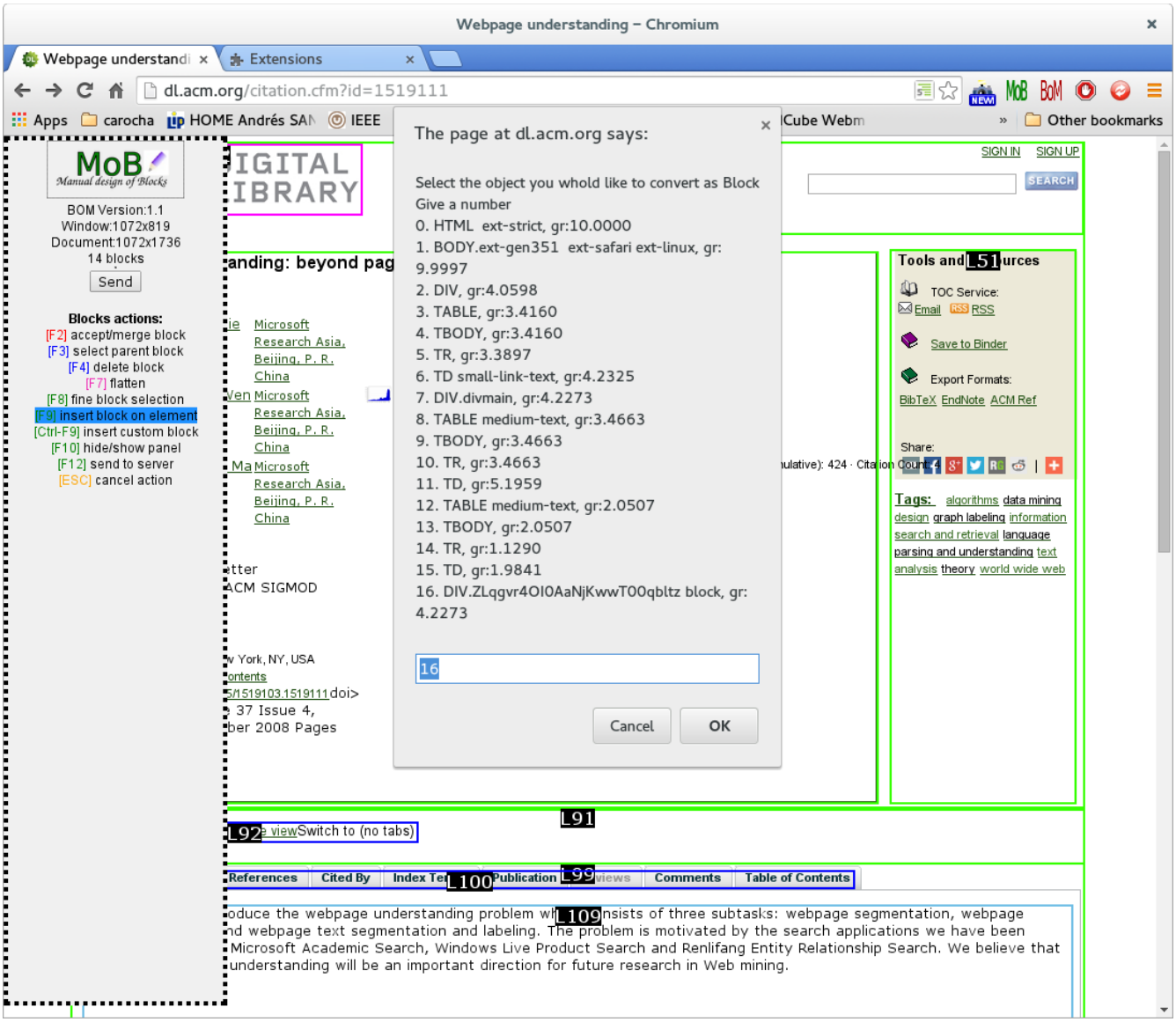
Fine block selection

Sometimes because of the hierarchy, one block can be behind its children. If we would like to access blocks in this situation we can use the F8 key. A simple menu will prompt to write the position number of the block on the list we want to select.



Insert block on element

If the tool does not propose or does not detect an element in the web page, we can create a new one with the F9 function. Clicking with the mouse in some area of the page will raise a prompt list of all the DOM elements under the click. We indicate the number of that we would like to create a block over.



Insert custom block

On the case the insert block by an element fails we can “draw” a custom block in the web page. This option has the aggravating that it does not count the underlying HTML elements and the word cover. It only “draws” the block.

There are needed two steps: first we mark the first point with the mouse and then the second point. A rectangle is drawn from point1 to point2.

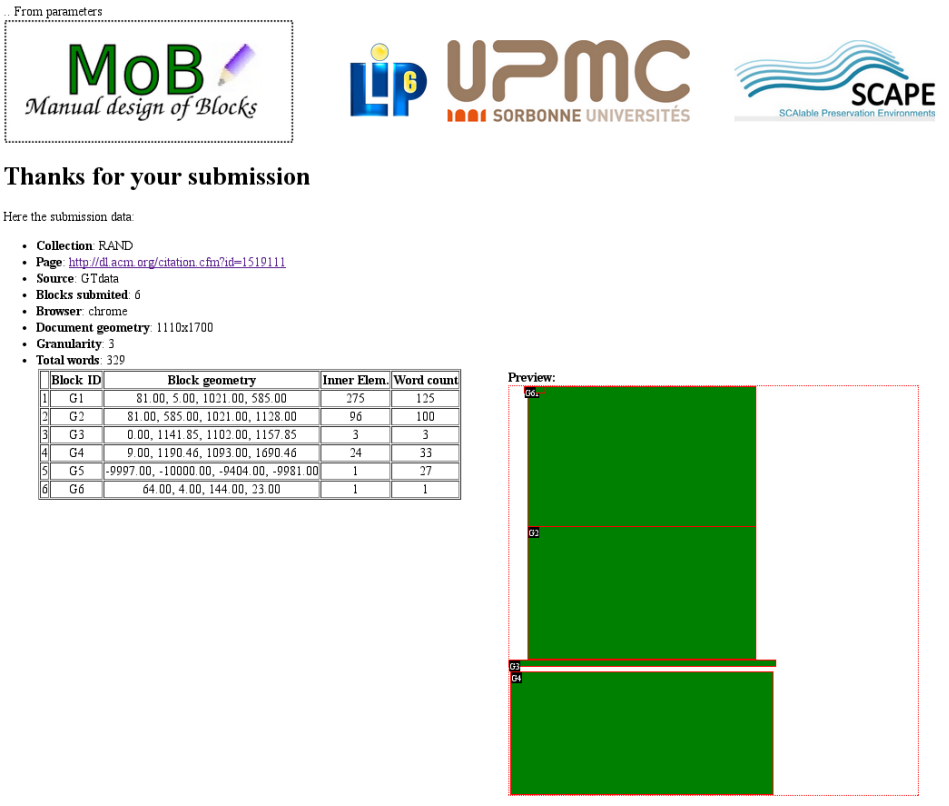
Sending to server

We evaluate only flat segmentations. We have to be careful that there are no hierarchy. If not sure we can press F7 flatten the segmentation.

To send the segmentation:

1. we click over the “Send” button in the tool bar.
2. Select the corresponding category (blogs, forums, etc)
3. Select the collection (usually GOSH is already selected)
4. Click over “Yes send to server”

A web page will appear confirming the submission. For example the following figure:



From parameters

MoB Manual design of Blocks

LIP6 UPMC SORBONNE UNIVERSITÉS SCAPE SCAlable Preservation Environments

Thanks for your submission

Here the submission data:

- **Collection**: RAND
- **Page**: http://dl.acm.org/citation.cfm?id=1519111
- **Source**: GTdata
- **Blocks submited**: 6
- **Browser**: chrome
- **Document geometry**: 1110x1700
- **Granularity**: 3
- **Total words**: 329

	Block ID	Block geometry	Inner Elem.	Word count
1	G1	81.00, 5.00, 1021.00, 585.00	275	125
2	G2	81.00, 585.00, 1021.00, 1128.00	96	100
3	G3	0.00, 1141.85, 1102.00, 1157.85	3	3
4	G4	9.00, 1190.46, 1093.00, 1690.46	24	33
5	G5	-9997.00, -10000.00, -9404.00, -9981.00	1	27
6	G6	64.00, 4.00, 144.00, 23.00	1	1

Preview:

Repository

All submission go to a repository that can be viewed in:

http://www-poleia.lip6.fr/~sanojaa/BOM/internal/inventory/

login: scape, password: scape2013

Blogs
http://hackerluddite.wordpress.com/2014/05/05/how-to-specify-the-firefox-binary-for-selenium-webdriver-under-node-js/
http://twitterthecomic.tumblr.com/
https://medium.com/teaching-learning/the-restaurant-and-the-kitchen-208274e15e0d
http://j2kun.svbtle.com/why-dont-researchers-write-great-code
http://cutepuppylove.me/2014/09/08/this-soldiers-dog-rescues-him-by-doing-something-i-never-knew-dogs-could-do/
http://bellacaledonia.org.uk/2014/09/08/labour-pains-labour-of-love/
http://wehuntedthemammoth.com/2014/09/08/zoe-quinns-screenshots-of-4chans-dirty-tricks-were-just-the-appetizer-heres-the-first-course-of-the-dinner-directly-from-the-irc-log/
http://willthef1journo.wordpress.com/2014/09/08/enough-is-enough/
http://remolacha.net/2014/09/la-nueva-presentadora-fui-fuiu-de-divertido-con-jochy.html
http://shariaunveiled.wordpress.com/2014/09/07/king-abdullah-of-saudi-arabia-issues-warning-isis-will-be-in-europe-in-one-month-and-us-in-two-months-video-report/
http://familyguyaddicts.com/2014/09/08/comic-con-wrap-up/
http://blogdopaulinho.wordpress.com/2014/09/08/gilmar-rinaldi-tem-atitude-deploravel-em-corte-de-maicon/
http://brokensilenze.net/2014/09/08/love-and-hip-hop-atlanta-season-3-episode-20-reunion-part-3-full-episode/
http://tanketornet.wordpress.com/2014/09/07/hej-alla-som-rostade-pa-fi-men-glomde-lasa-partiprogrammet/
http://koreanindo.net/2014/09/08/polaris-entertainment-umumkan-tempat-peristirahatan-terakhir-untuk-rise/
http://cahidejibek.com/2014/09/08/aci-sos-konservesi/
http://anotherscotland.wordpress.com/2014/09/07/a-letter-to-england/
http://jessicasmmnds.wordpress.com/2014/09/08/to-ryan-kylie-murphy/
http://jamstalldhetsfeministern.wordpress.com/2014/09/06/rosta-pa-feministiskt-initiativ-for-karlekens-skull/
http://hans.quora.com/
Forum (threads)
http://forums.marvelheroes.com/discussion/127964/this-game-needs-a-way-to-kick-inactive-players-in-x-def
http://placeshiftingenthusiasts.com/forum/slingplayer-for-connected-devices/roku-and-slingplayer-connection-question/
https://forum.videolan.org/viewtopic.php?f=14&t=79953
http://talkiforum.com/forum-example/#/20110213/i-post-on-your-forum-you-post-on-mine-any-tho-364923/
http://www.theflirtingshack2.com/showthread.php?10821-Duplicate-User-Accounts
http://www.swapitforum.co.uk/showthread.php?31873-Complete-Forum-Beginners-Guide!
http://www.flirtsnfriends.com/forum/married-flirting-resort/8-hello-flirty-friends.html
http://discussionlounge.co.uk/viewtopic.php?f=2&t=413

http://prisonbreakforum.co.uk/Heroes_Back_to_the_Day_Job_about5330.html
http://www.wrestlingsmarks.com/threads/suggestion-thread.73369/
http://www.thewweforum.com/threads/official-wwe-network-thread-questions-concerns-discussion.4749/
http://www.occasion-to-be.com/forum/general-discussion/fifth-viking-'ring-fortress'-discovered-in-denmark-may-have-been-military-traini/
http://z11.invisionfree.com/After_Sundown/index.php?showtopic=10
http://cryptx.org/news-and-announcements/32-cryptx-coin-shop.html
http://wowsucks.freeforums.org/recruiting-contest-t133.html
http://durellia.proboards.com/thread/1392/adult-baby-source-movie-pack
http://www.lilwaynehq.com/forums/introduce-yaself/28-how-did-you-find-out-about-forum.html
http://s3.excoboard.com/Pure_Fun/35189/196383
http://www.onlinedebate.net/forums/showthread.php/27177-Memorial-Day?s=56aa0b697b3a62e9690987534f99dc1c
http://www.mojo-themes.com/forums/topic/invoice-3/
Wiki
http://iclassroom.pbworks.com/w/page/12663802/Home
https://zohoshow.wiki.zoho.com/
http://community.wikia.com/wiki/User_blog:Sarah_Manley/Community_Contest_-_May_the_best_wiki_win#comm-245895
http://ilek.wikispot.org/Freizeitkarte
http://www.aboutus.org/Spring-Valley-NV-United-States
http://www.appropedia.org/MY4777
http://baike.baidu.com/subview/117228/5039325.htm
http://ballotpedia.org/Alaska
http://www.bullshit.wiki/A_List_of_Pointless_Websites
http://en.citizendium.org/wiki/Astronomy
http://veiling.catawiki.nl/kavels/740293-miniatuur-telecaster-gitaar
http://www.prwatch.org/node/12586
http://www.conservapedia.com/Atheist_cults
http://connectipedia.org/Artists
http://www1.cpdl.org/wiki/index.php/Alexander_Dmitriyevich_Kastalsky
http://daviswiki.org/Downtown
http://www.ecured.cu/index.php/Batalla_de_Waterloo
http://enciclopedia.us.es/index.php/Idioma_espa%C3%B1ol
http://fr.ekopedia.org/Gonfler_un_pneu
http://www.fringepedia.net/wiki/Krista_Manning

http://www.thebridalstudiosouthport.co.uk/
http://www.galaxiki.org/cgi-bin/local/perl/galaxiki/wiki.pl?mode=star&id=1413927
http://www.foodista.com/blog/2014/09/06/copycat-asian-takeout-recipes
http://www.gardenology.org/wiki/Albizzia_stipulata
http://www.geonames.org/countries/VE/venezuela.html
http://heroeswiki.com/Lesli_Glatter
http://www.geo-wiki.org/
http://en.memory-alpha.org/wiki/Luyten_789-6
http://lgbthistoryuk.org/wiki/index.php?title=George_Frideric_Handel
http://imslp.org/wiki/Keyboard_Sonata_No.3_in_G_minor_(Bersanetti,_Gianluca)
http://meatballwiki.org/wiki/IsmellDevice
http://lostpedia.wikia.com/wiki/Oahu
http://hlwiki.slais.ubc.ca/index.php/Bibliographic_citation_software
http://lyrics.wikia.com/Nicki_Minaj:Anaconda
http://www.baike.com/wiki/%E7%BE%8E%E5%9B%BD&prd=shouye_newslist
http://metadatabase.org/wiki/ValLigURL
http://en.metapedia.org/wiki/Acad%C3%A9mie_fran%C3%A7aise
http://mywikibiz.com/Develop_A_Second_Income
http://openwetware.org/wiki/Julius_B._Lucks/Projects/Python_All_A_Scientist_Needs
http://pcgamingwiki.com/wiki/Full_Bore
http://rywiki.tsadra.org/index.php/blo_rigs
http://www.encyclopediaofmath.org/index.php/Green_equivalence_relations
Picture
http://www.shutterfly.com/photo-book/storytelling/archives/ornaments-for-all-seasons--festive-fall-wreath.sfly
http://tinypic.com/search.php?tag=earth+day&type=images#.VAmxZDRQTRZ
https://www.imageshack.us/discover
http://instagram.com/mbfashionweek? utm_source=partner&utm_medium=embed&utm_campaign=photo
http://www.mnn.com/lifestyle/eco-tourism/stories/too-beautiful-to-be-real-16-surreal-landscapes-found-on-earth
http://instagram.com/marcjacobsintl? utm_source=partner&utm_medium=embed&utm_campaign=photo
http://instagram.com/tibi?utm_source=partner&utm_medium=embed&utm_campaign=photo
http://instagram.com/zac_posen?utm_source=partner&utm_medium=embed&utm_campaign=photo
http://imgur.com/gallery/EoIvS
http://imgur.com/gallery/stXI7ed
http://twitpic.com/ea06o2

http://twitpic.com/e9qgmf
http://www.snapfish.com.au/snapfishau/photobook
Enterprise (corporate home page portals)
http://www.bp.com/
http://www.eni.com/en_IT/home.html
http://www.sap.com/index.html
www.siemens.com/
http://www.chevron.com/
http://www.cisco.com/
http://www.loreal.com/default.aspx
http://www.bnpparibas.com/
http://www.hsbc.com/
http://www.total.com/en/
http://corporate.walmart.com/
http://www.vale.com/brasil/EN/Pages/default.aspx
http://www.shell.com/

Metodology
Blogs pages are crawled from the Blog top post indexes.
Forums are crawled from Forum web indexes randomly.
Picture are crawled from top picture post to follow by each site.
Enterpireses where crawled from bowen craggs index