

**Organization Agreement  
to use the**

**ClueWeb09 Web Research Collections**

The \_\_\_\_\_ (“Group”), a group or division of approximately \_\_\_\_\_ people engaging in research and development of natural-language-processing, information-retrieval, or document-understanding systems, is part of the following corporation/partnership/legal entity listed below (the “Organization”).

Corporation/Partnership/Legal Entity \_\_\_\_\_

Official mail address \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Contact name \_\_\_\_\_

Telephone \_\_\_\_\_

Facsimile \_\_\_\_\_

The Group would like to use the information designated as the ClueWeb09 Text Research Collections (the "Information"). By signing this Organization Agreement (“Agreement”) with Carnegie Mellon University (“Carnegie Mellon”), the Organization hereby agrees to abide by the following understandings, terms and conditions. These understandings, terms and conditions apply equally to all or to part of the Information, including any updates or new versions of the Information supplied under this Agreement. Organization understands and agrees that the licenses to the Information granted under this Agreement are for use only by persons working within Organization’s specific Group identified above, subject to the terms and conditions below.

**Copyright**

1. The Information has been obtained by crawling the Internet. Due to the amount of Information it has not been practicable to obtain permission from copyright owners to provide the Information for the uses permitted under this Agreement (“Permitted Uses”).
2. Organization understands that all the documents in the Information are documents which have been at some time made publicly available on the Internet and which have been collected using a process which respects the commonly accepted methods (such as robots.txt) for indicating that the documents should not be so collected.
3. Owners of copyright in individual documents may choose to request deletion of these documents from the Information.

4. The limitation on permitted use contained in the following section is intended to reduce the risk of any action being brought by copyright owners, but if this happens the Organization agrees to bear all associated liability.

### **Permitted Uses**

1. The Information may only be used for research and development of natural-language-processing, information-retrieval or document-understanding systems by the specific Group identified above.
2. Summaries, analyses and interpretations of the linguistic properties of the Information may be derived and published, provided it is not possible to reconstruct the Information from these summaries.
3. Small excerpts of the Information may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development carried out and related issues.
4. All efforts must be made not to infringe the rights of any third party including, but not limited to, the authors and publishers of any excerpts used in accordance with clause 3 above in this “Permitted Uses” section.

### **Own Assessment of Information Use**

The Organization must make its own assessment of the suitability of the Information for its research and development purposes under Permitted Uses.

### **Agreement to Delete Data on Request**

The Organization shall delete within thirty days of receiving notice all copies of any particular document that is part of the Information whenever requested to do so by either:

1. Carnegie Mellon; or
2. the owner of copyright for the particular document.

### **Access to the Information by Individuals**

The Organization:

1. must control access to the Information by individuals and may only grant access to people within the identified Group who are working under its control, i.e., its own employees, consultants under written agreement to the Organization, or individuals providing service to the Organization under written agreement;
2. must ensure that before being given access an individual must complete and submit the Individual Agreement form;

3. must terminate an Individual's access when the individual no longer requires access for its work for the Organization and/or no longer is employed by (and/or under contract with, as applicable) the Organization;
4. remains responsible for any breach of the Individual Agreement form by individuals to whom Organization has granted access to the Information;
5. shall retain the applications of all persons ever granted access to the Information and make them available upon request to any of the copyright holders and to Carnegie Mellon;
6. shall maintain a list of people with current and recently-terminated access to the Information and make it available to Carnegie Mellon on request; and
7. must make sure that an Individual with access displays the Information to or shares the Information with only persons whom his or her Organization lists as having access to the Information.

### **No Warranty; Disclaimers**

NO WARRANTY IS GIVEN BY CARNEGIE MELLON IN RELATION TO THE INFORMATION OR THE USE(S) TO WHICH THE INFORMATION MAY BE PUT BY THE ORGANIZATION OR THE FITNESS OR SUITABILITY OF THE INFORMATION FOR ANY PARTICULAR PURPOSE OR UNDER ANY SPECIAL CONDITIONS REGARDLESS OF WHETHER ANY SUCH PARTICULAR PURPOSE OR SPECIAL CONDITIONS ARE OR HAVE BEEN MADE KNOWN TO CARNEGIE MELLON PRIOR TO, OR DURING, THE PERIOD OF THIS AGREEMENT. ALL CONDITIONS AND WARRANTIES EXPRESS OR IMPLIED, WHETHER ARISING UNDER STATUTE OR UNDER COMMON LAW, INCLUDING BUT NOT LIMITED TO CONDITIONS AND WARRANTIES AS TO QUALITY, MERCHANTABILITY AND FITNESS FOR PURPOSE ARE HEREBY EXCLUDED. CARNEGIE MELLON SHALL NOT BE LIABLE TO ORGANIZATION OR ANY THIRD PARTY FOR LOSS OF PROFITS OR FOR INCIDENTAL, INDIRECT, SPECIAL OR CONSEQUENTIAL DAMAGES FOR ANY REASON WHATSOEVER ARISING OUT OF OR RELATING TO THIS AGREEMENT (INCLUDING ANY BREACH OF THIS AGREEMENT), EVEN IF CARNEGIE MELLON HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES OR HAS OR GAINS KNOWLEDGE OF THE EXISTENCE OF SUCH DAMAGES.

### **Termination**

Either party may terminate this Agreement at any time by notifying the other party in writing. On termination, the Organization must: a) immediately cease using the Information; and b) return to Carnegie Mellon or destroy all copies of the Information and notify Carnegie Mellon in writing of the action taken.

### **Applicable Law; Disputes**

This Agreement is governed by the laws of the Commonwealth of Pennsylvania in the United States of America. All claims and/or controversies of every kind and nature arising out of or relating to this Agreement, including any questions concerning its existence, negotiation, validity,

meaning, performance, non-performance, breach, continuance or termination shall be settled exclusively in the United States District Court for the Western District of Pennsylvania or, if such Court does not have jurisdiction, in any court of general jurisdiction in Allegheny County, Pennsylvania and each party consents to the exclusive jurisdiction of any such courts and waives any objection which such party may have to the laying of venue in any such courts.

### **Notices**

Any notice to the other party under this Agreement must be in writing, signed by the party giving it, and provided either personally, by registered mail, certified mail, and/or reputable overnight courier (such as UPS, FedEx, etc.). Notices to Organization shall be sent using the contact person and address listed on the initial page of this Agreement. Notices to Carnegie Mellon shall be sent to the following person and address:

Associate Vice President, Research  
Carnegie Mellon University  
5000 Forbes Ave.  
WQED Building  
Pittsburgh, PA 15213  
USA

telephone: +1 412-268-8746

fax: +1 412-268-6279

Either party may update its contact information by providing written notice to the other party as required by this Section.

### **Miscellaneous**

If any portion of this Agreement is determined by any court or governmental agency of competent jurisdiction to violate applicable law or otherwise not to conform to requirements of law, then the rest of the Agreement will remain in effect and the parties will substitute a suitable and equitable provision for the invalid/unenforceable provision in order to carry out the original intent and purpose of the original Agreement. Organization may not assign any or all of its rights and/or obligations under this Agreement without the prior written consent of Carnegie Mellon, which consent may be granted or withheld in Carnegie Mellon's sole discretion. Any attempted assignment in violation of this section shall be void and of no effect. This Agreement constitutes the entire agreement between the parties and supersedes all previous agreements and understandings relating to the subject matter of this Agreement. The Agreement may not be altered, amended, or modified except by a written instrument signed by the duly authorized representatives of both parties.

Intending to be legally bound, Organization and Carnegie Mellon execute this Agreement effective as of the date the last party signs.

**By the Organization:**

*By signing below, I represent and warrant that I have authority to bind the Organization to the terms of this Agreement*

Signature \_\_\_\_\_

Date \_\_\_\_\_

Name (please print) \_\_\_\_\_

Title \_\_\_\_\_

**Accepted by Carnegie Mellon University:**

Signature \_\_\_\_\_

Date \_\_\_\_\_

Name (please print) \_\_\_\_\_

Title \_\_\_\_\_

***Please complete, this form and return it to the address below:***

TO: Carnegie Mellon University  
Web Test Collections  
EMAIL: [clueweb@cpftgy.cmu.edu](mailto:clueweb@cpftgy.cmu.edu)

## ClueWeb09 Order Form

1. **Which version of the dataset do you want?** (Pick just one.)

- Dataset license only (cost is \$0):
- The full dataset (1 × 8 TB disk) (cost is \$380):
- The TREC “Category B” subset (1 × 500 GB disk) (cost is \$185):
- The NTCIR Intent Task subset (1 × 500 BB disk) (cost is \$185):
- The TREC 2011 Crowdsourcing subset (download) (cost is \$0):

2. **How would you like to receive an invoice (if one is required)?**

- Paper + postal service (slow)
- Pdf + email (fast)

3. **Where should the invoice be sent (if one is required)?** (Please provide both mail & email)

Name: \_\_\_\_\_

Organization: \_\_\_\_\_

Mailing address: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Email address: \_\_\_\_\_

4. **Payment method (if required):** Credit Card  Check  Wire transfer   
(If you select the credit card option, we will contact you to obtain your credit card number)

5. **Where should the dataset be shipped (if shipping is required)?**

Name: \_\_\_\_\_

Organization: \_\_\_\_\_

Mailing address: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Telephone: \_\_\_\_\_

6. **Preferred shipping method (if required):** 1 day  3 day  1 week

7. **Returning the agreement and order form:** Please return the completed form as a pdf file to clueweb at andrew dot cmu dot edu.