TREC 2007 Genomics Track Guidelines for Relevance Judges

The goal of the TREC Genomics Track is to improve information retrieval in the biomedical domain. Each year, 30-40 research groups develop algorithms that automatically interpret questions collected from real biologists into queries that can be applied against databases of scientific articles. The groups submit their best search results, which are pooled and judged for relevance to the original question. This year, 160,000 full-text articles from 59 journals will be searched. See http://ir.ohsu.edu/genomics/2007protocol.html for more details.

The task for this year has 36 questions that are answered by specific types of information, or entities, such as genes, proteins, mutations, or disease. Your job as a relevance judge is to assess the relevance of submitted paragraphs, isolate the minimum information in the paragraph that answers the question, and group relevant answers by similarity.

This document is divided into three major sections. Part I offers guidelines for making your relevance judgments and selecting entities that answer the question in a consistent, repeatable manner. Part II covers the Relevance Judging Database, including how to install it, enter your judgments and entities, save results, and return them to the organizers. Part III has contact information.

Thanks again for your participation.

Part I: Relevance Assessments, Minimal Excerpts, and Selecting Entities

1. Review the question and identify key concepts

   Upon receipt of your topic, first review the question. If there is a gene or protein mentioned, identify synonyms for it. For biological processes or diseases, familiarize yourself with more general concepts, as well as sub-topics (see Appendix I for suggested resources). For example, “mad cow disease” in the 2006 Topic 160 is formally known as bovine spongiform encephalopathy, abbreviated as BSE. It is a member of the Transmissable Spongiform Encephalopathies (TSE) disease family, of which Creutzfeld-Jacob disease (CJD) is another member (these relationships may be found by looking up the disease names in MeSH, listed in Appendix I). Therefore, definitely relevant passages refer to mad cow or BSE, possibly relevant passages refer to the TSE family, of which mad cow disease is a member. References to the related, but different disease, CJD, are not relevant. In another example, Topic 179 asks about liver function. The definition of “liver development” in the Gene Ontology (also listed in Appendix I) states that the liver secretes bile, synthesizes blood clotting factors and vitamin A, and stores glycogen. Those functions that are unique to the liver and are supplied as an answer are definitely relevant. References to functions that occur in other organs in addition to the liver are possibly relevant.

2. Identify relevant paragraphs and select minimum complete excerpts.

   You will receive an OpenOffice.org database file that contains paragraphs from full text journal articles. The paragraphs were identified by automated search and retrieval algorithms as part of the 2007 TREC Genomics track. Your first task is to determine whether the paragraphs contain complete answers to the topic questions. .Table 1 shows examples of judgments from the 2007 training topics and the reasoning behind the

judgments. In general, a paragraph is ***definitely relevant*** if it contains all key concepts of the question AND it has the required entity that answers the question. A paragraph is ***possibly relevant*** if it contains the majority of key concepts, if missing concepts are within the realm of possibility (i.e. more general terms are mentioned that probably include the missing concepts), AND it has the required entity that possibly answers the question.

If the paragraph is definitely relevant or possibly relevant, you will select the minimum amount of text (called Answer Text) that answers the question completely. The Answer Text MUST contain a member of the entity class stated in the question. Pronouns (e.g., *these*, *they*, *it*) that do not reference the relevant subject within the extracted text are inadequate. Reference to the relevant subject as an unspecified generic component (e.g., "the subunit", "these genes", "this disease") defined within the extracted text is acceptable ONLY if they are defined elsewhere within the Answer Text extract. Acronyms or abbreviations that are explained outside the Answer Text but whose definitions are not part of the excerpt are acceptable, as these are considered to be functioning as synonyms. Minimum excerpts may range from a portion of a sentence to the entire paragraph, but they must contain all key concepts from the question to be definitely relevant AND the entity term that answers the question.

Table 1. Examples and Reasoning for Relevance Judgments. Entity terms are underlined.

Topic ID	Question	Excerpt	Judgment	Reason
T10	What [PROTEINS] does epsin1 interact with during endocytosis?	We have proposed that on binding to membranes, this new helix buries itself between the lipid headgroups, pushing the lipids apart and thus promoting membrane curvature in the clathrin-coated bud (Ford et al., 2002). As this will need to be a coordinated event, the COOH terminus of epsin1 binds to both clathrin and the AP2 complex, thus inducing the curvature in newly forming coated pits	Definitely Relevant	Even though endocytosis isn't explicitly mentioned, clathrin-coated bud formation is part of the endocytic process.
T10	What [PROTEINS] does epsin1 interact with during endocytosis?	Epsin1 is a cytosolic protein required for endocytosis. An additional pool of epsin1 is present in the nucleus in a complex with the transcription factor PLZF (28). Tubulin interacts with all ENTH and ANTH domains tested (epsin1, epsinR, AP180, HIP1, and Hip1R)."	Possibly Relevant	Epsin1-interacting proteins may be involved in endocytosis, but it is not explicitly stated.
T10	What [PROTEINS] does epsin1 interact with during endocytosis?	These motifs show a great degree of sequence similarity to the N-terminal sequence of Epsin1, a protein that contains additional binding regions that interact with ubiquitin, clathrin and the ear domain of AP-2 (Ford et al., 2002). ENTH domains (dubbed from epsin N-terminal homology) have since been recognised in several proteins that participate in clathrin-mediated endocytosis or	Possibly Relevant	Epsin1-interacting proteins may be involved in endocytosis, but it is not explicitly stated.

		vesicle budding (Legendre-Guillemin et al., 2004).		
T6	What centrosomal [GENES] are implicated in diseases of brain development?	Fibroblast lines derived from R6/2 mice and from HD patients were found to have a high frequency of multiple centrosomes which could account for all of the observed phenotypes including a reduced mitotic index, high frequency of aneuploidy and persistence of the midbody We have previously generated the R6/2 mouse model that expresses exon 1 of the human HD gene containing CAG repeats in excess of 150..	Possibly Relevant	Strictly speaking, a centrosomal gene is one that encodes a centrosomal subunit, but a looser definition would include genes that regulate centrosome biogenesis.
T6	What centrosomal [GENES] are implicated in diseases of brain development?	We found that virtually all Brca1 11/ 11Gadd45a-/- embryos exhibited exencephaly, showing increased apoptosis in their neuroepithelia due to p53 activation, as haploid or complete loss of p53 repressed apoptosis and rescued embryonic lethality. Our further analysis uncovered a synergistic role of Brca1 and Gadd45a in regulating centrosome duplication and in maintaining genome integrity.	Possibly Relevant	Strictly speaking, a centrosomal gene is one that encodes a centrosomal subunit, but a looser definition would include genes that regulate centrosome biogenesis.
T6	What centrosomal [GENES] are implicated in diseases of brain development?	In order to determine how mutations in DISC1 might cause susceptibility to schizophrenia, we undertook a comprehensive study of the cellular biology of DISC1 in its full-length and disease-associated mutant forms. DISC1 interacts by yeast two-hybrid, mammalian two-hybrid, and co-immunoprecipitation assays with multiple proteins of the centrosome and cytoskeletal system, including MIPT3, MAP1A and NUDEL	Definitely relevant	DISC1 is associated with schizophrenia and co-localizes with known centrosomal subunits.

T6	What centrosomal [GENES] are implicated in diseases of brain development?	The hallmark of the 8p12 stem cell myeloproliferative disorder (MPD) is the disruption of the FGFR1 gene, which encodes a tyrosine kinase receptor for members of the fibroblast growth factor family… We report here the cloning of the t(8;9)(p12;q33) and the detection of a novel fusion betweenFGFR1 and the CEP110 gene, which codes for a novel centrosome-associated protein with a unique cell-cycle distribution	Not Relevant	Myeloproliferative disorder is not a brain development disease.
T8	What [MUTATIONS] in apolipoproteins are associated with disease?	To determine the frequency of familial hypoalphalipoproteinemia in the general population due to mutation of the apolipoprotein A-I (apo A-I) gene, we analyzed sequence variations in the apo A-I gene	Not Relevant	Even though a disease and an apolipoprotein is mentioned, a specific mutation is not.
T8	What [MUTATIONS] in apolipoproteins are associated with disease?	Using single strand conformation polymorphism (SSCP) analysis and followed by sequencing of DNA amplified from the 67 individuals with low HDL-C levels, we identified mutations in five subjects in the heterozygous state with the wild-type allele (Fig. 2). Three were frameshift mutations. The first frameshift mutation was a single C nucleotide insertion in codons 3-5 where seven consecutive C residues are found.	Possibly Relevant	Low HDL-C levels is not a disease *per se*, but it is predictive of coronary heart disease.
T1	What [ANTIBODIES] have been used to detect protein TLR4?	An aliquot of cytoplasmic protein (20–100 μg) was utilized for Western blotting with specific primary antibodies (Santa Cruz Biotechnology) to TLR4 (sc-10741)	Definitely Relevant	Santa Cruz Biotech. makes primary antibodies to TLR4 called “sc-10741”.
T1	What [ANTIBODIES] have been used to detect protein TLR4?	The goat anti–TLR4 antibody, the rabbit anti–MyD88 antibody, the mouse anti–IL-1R–associated kinase (IRAK)-1 antibody, as well as the rabbit anti–p65 antibody were purchased from Santa Cruz Biotechnology	Not relevant	The name of the anti-TLR4 antibody from Santa Cruz Biotech. Is not stated.
T1	What [ANTIBODIES] have been used to detect protein TLR4?	anti-TLR4 mAb (HTA1216, a gift from Dr. Kensuke Miyake, University of Tokyo, Tokyo, Japan).”	Definitely Relevant	An antibody against TLR4 called “HTA1216” was provided by Dr. K. Miyake.
T1	What [ANTIBODIES] have been used to detect protein TLR4?	Murine anti-human TLR4 monoclonal antibodies were from Dr. K. Miyake	Not relevant	The name of antibody provided by Dr. Miyake is not stated.

IMPORTANT! There are some questions for which there is little published information. It is tempting to show leniency when few relevant paragraphs are encountered. Resist the urge to relax criteria for relevance and try to maintain consistent evaluation standards. One way to guard against "relevance drift" is to review your judgments made early in the process after you have completed a topic.

3. Develop controlled vocabularies and code results

Relevant Answer Text excerpts that come from different articles may contain largely the same information (e.g. research that is frequently cited in the introduction of subsequent articles). Participants in TREC 2007 will be rewarded for the breadth of answers submitted. To determine breadth, excerpts will be coded with standardized terms, selected by you, allowing grouping of related results.

Table 2 shows the entity types for which you will develop controlled vocabularies. Each topic is answered by an entity, which will be found in the excerpt of text that you selected in step 2. Once you have selected relevant Answer Text, you will identify entity terms and collapse all synonymous terms into a single Primary Entity Term. Examples of synonymous terms are shown in Table 4.

Table 2. Entity Types and Definitions

These entities are based on controlled terminologies from different sources, with the source of the terms depending on the entity type. Below is a table of the entity types. See Appendix I for links to the suggested term sources.

Entity Type	Definition	Potential Source of Terms
ANTIBODIES	Immunoglobulin molecules having a specific amino acid sequence by virtue of which they interact only with the antigen (or a very similar shape) that induced their synthesis in cells of the lymphoid series (especially plasma cells).	
BIOLOGICAL SUBSTANCES	Chemical compounds that are produced by a living organism.	MeSH
CELL OR TISSUE TYPES	A distinct morphological or functional form of cell, or the name of a collection of interconnected cells that perform a similar function within an organism.	MeSH
DISEASES	A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown.	MeSH
DRUGS	A pharmaceutical preparation intended for human or veterinary use.	Medline Plus
GENES	Specific sequences of nucleotides along a molecule of DNA	iHoP,

	(or, in the case of some viruses, RNA) which represent functional units of heredity.	Harvester
MOLECULAR FUNCTIONS	Elemental activities, such as catalysis or binding, describing the actions of a gene product or bioactive substance at the molecular level.	AmiGO browser
MUTATIONS	Any detectable and heritable change in the genetic material that causes a change in the genotype and which is transmitted to daughter cells and to succeeding generations	
PATHWAYS	A series of biochemical reactions occurring within a cell to modify a chemical substance or transduce an extracellular signal.	BioCarta, KEGG
PROTEINS	Linear polypeptides that are synthesized on ribosomes and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits.	iHoP, Harvester
STRAINS	A genetic subtype or variant of a virus or bacterium.	
SIGNS OR SYMPTOMS	A sensation or subjective change in health function experienced by a patient, or an objective indication of some medical fact or quality that is detected by a physician during a physical examination of a patient.	MeSH
TOXICITIES	A measure of the degree and the manner in which which something is toxic or poisonous to a living organism.	MeSH
TUMOR TYPES	An abnormal growth of tissue, originating from a specific tissue of origin or cell type, and having defined characteristic properties, such as a recognized histology.	MeSH

Part II: How to Use the Relevance Judgment Tool

In general, ***save your work after every change***.

Download Open Office from http://download.openoffice.org/2.2.1/index.html.

OpenOffice.org is cross-platform and runs on Windows, OS X, and Linux. Follow the instructions corresponding to your operating system.

1. Start the OpenOffice.org Base (database) program by opening (double-click or choose open from the file menu) the .odb file for your topic. Alternately, start the Open OpenOffice.org Base program and then open the .odb file from within Base.,
2. A file containing a database with the paragraphs to be judged will be emailed to you. To open your topic, from within Open Office, select File > Open > FileNameEndingIn.odb (Figure 1).
3. In Forms section, open both forms: ***Enter Relevance Judgments*** (Figure 2) and ***Add Entities Here*** (Figure 3).
4. Click on ***Data Source as Table*** icon in both forms (this will help with saving your work and picking up where you left off). If you can't find the icon, click the down-arrow at the

bottom right of the lower window button bar, select ***Visible Buttons***, and check off ***Data Source as Table***.

5. Read PLAIN TEXT and decide the level of relevance as explained above.
6. In the lower pane, use the RELEVANCE drop-down menu to enter your judgment.
7. If your judgment is Not Relevant, advance to the next record using the arrows on the bottom toolbar. SAVE YOUR WORK BY CLICKING ON THE SAVE ICON (looks like a little floppy disk).
8. If your judgment is Possibly or Definitely Relevant, enter your judgment. SAVE YOUR WORK.
9. Copy the minimum amount of text required to answer the question, and paste it into the lower pane ANSWER TEXT box. SAVE YOUR WORK. (Mac users: instead of copy and paste, you must drag and drop. Select the appropriate text from PLAIN TEXT and drag it to the ANSWER TEXT box.)
10. Copy the ENTITY term, go to the Add Entities Here form and paste it into the VALUES box (or type it in if that's easier). SAVE YOUR WORK. Use the NOTES section to add information that will help you collapse your entities into a minimal set of terms, such as the record number that prompted you to enter it, a definition, or additional information about the entity selection. SAVE YOUR WORK.
11. Advance to the next Entity record using the arrows on the bottom toolbar. CAUTION! Be sure the VALUE field is empty, or you will overwrite the previous entity.
12. When you have finished your relevance assessments, return to the ***Add Entities Here*** form. Reduce the entities to a minimal list, combining synonymous terms into one entity. The level of specificity is up to you, and it may require some research to determine whether entities should be grouped, especially with genes and proteins, which can go by several different names.
13. After reducing entities to a minimal set, and before adding entities to the Relevance Judgments, click the Refresh icon on both forms (this looks like a curved arrow). You must do this to see the final set of entities.
14. Add Entities to Relevance Judgments. Go to Enter Relevance Judgments form. Using the upper table view, advance to records judged Definitely and Possibly Relevant and make sure there is ANSWER TEXT, then add entities by selecting them from the dropdown menu. SAVE YOUR WORK after adding each entity. Add up to six entities to the Relevance Judgments, but make sure to fill the entity boxes in numeric order: ENTITY1, then ENTITY2, then ENTITY3, and so on. Every Definitely or Possibly Relevant passage must have at least ENTITY1 filled in. Every Not Relevant passage must leave ENTITY1 through ENTITY6 blank.

When you are finished, you will email the **.odb** file to Phoebe, Aaron and Bill. Before you send it, please check the following items:

1. Does every record have a relevance judgment?
2. Does every Definitely or Possibly Relevant record have ANSWER TEXT?
3. Does every Definitely or Possibly Relevant record have at least one ENTITY with the ENTITY1 drop down filled in?
4. Are there any duplicate entities?

To make this error checking step easier, there are reports that you can run which will display incorrect records. You can access these by choosing the Reports icon on the main window Database pane, and then double-clicking the report of your choice. You should run all of the reports before returning the filled-out judging form. Any record shown in a report is in error in

some way and needs to be fixed. A properly filled out form will have no records shown in any of these reports.

Congratulations! You're done! Email the .odb file and the **number of hours it took you to complete it** to phoebe.m.Roberts@gmail.com, cohenaa@ohsu.edu, hersh@ohsu.edu. You must include your hours in order to get paid.

Figure 1. The initial view upon opening the .odb file (numbers correspond to steps in Part II).

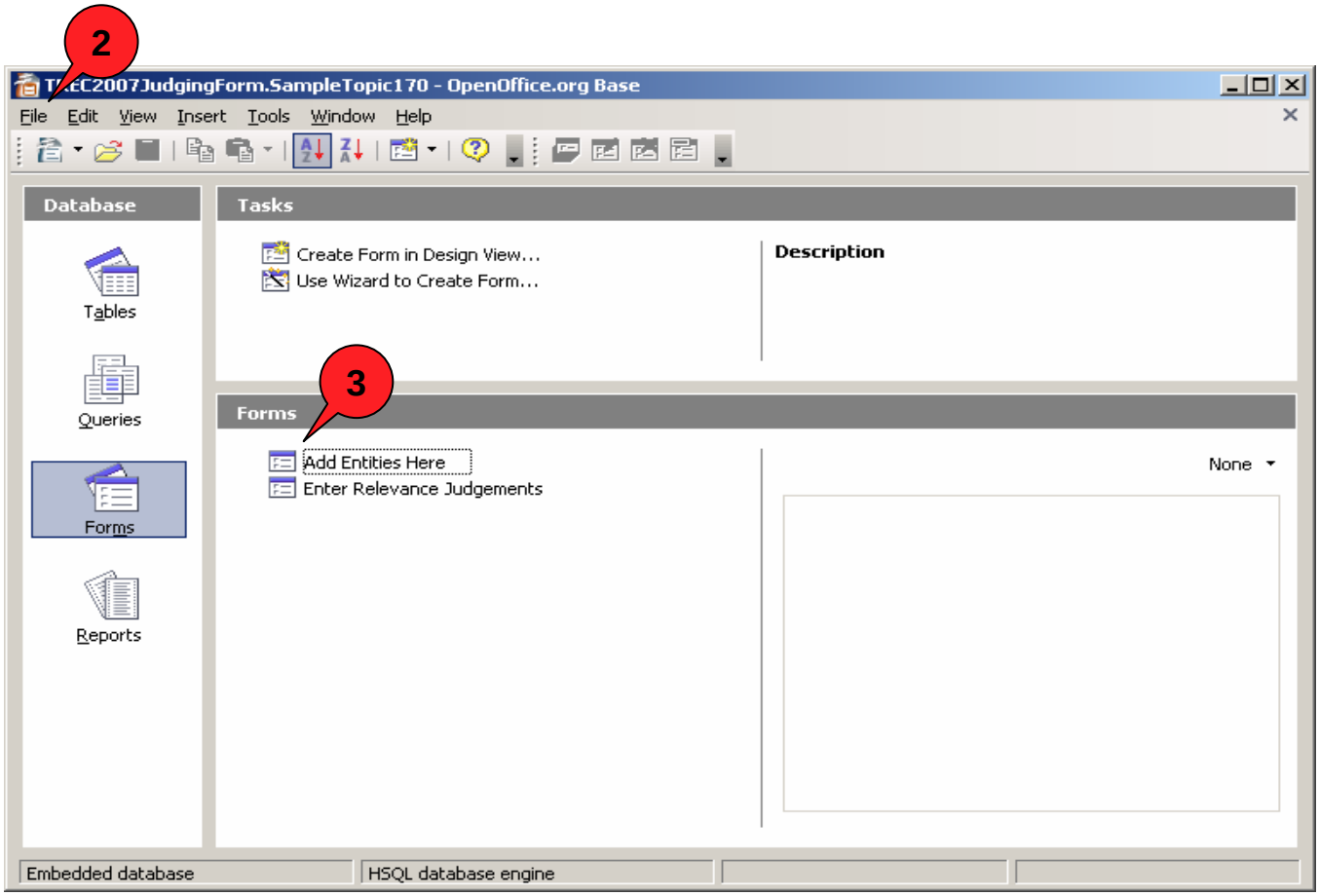


Figure 2. The Enter Relevance Judgements Form (numbers correspond to steps in Part II).

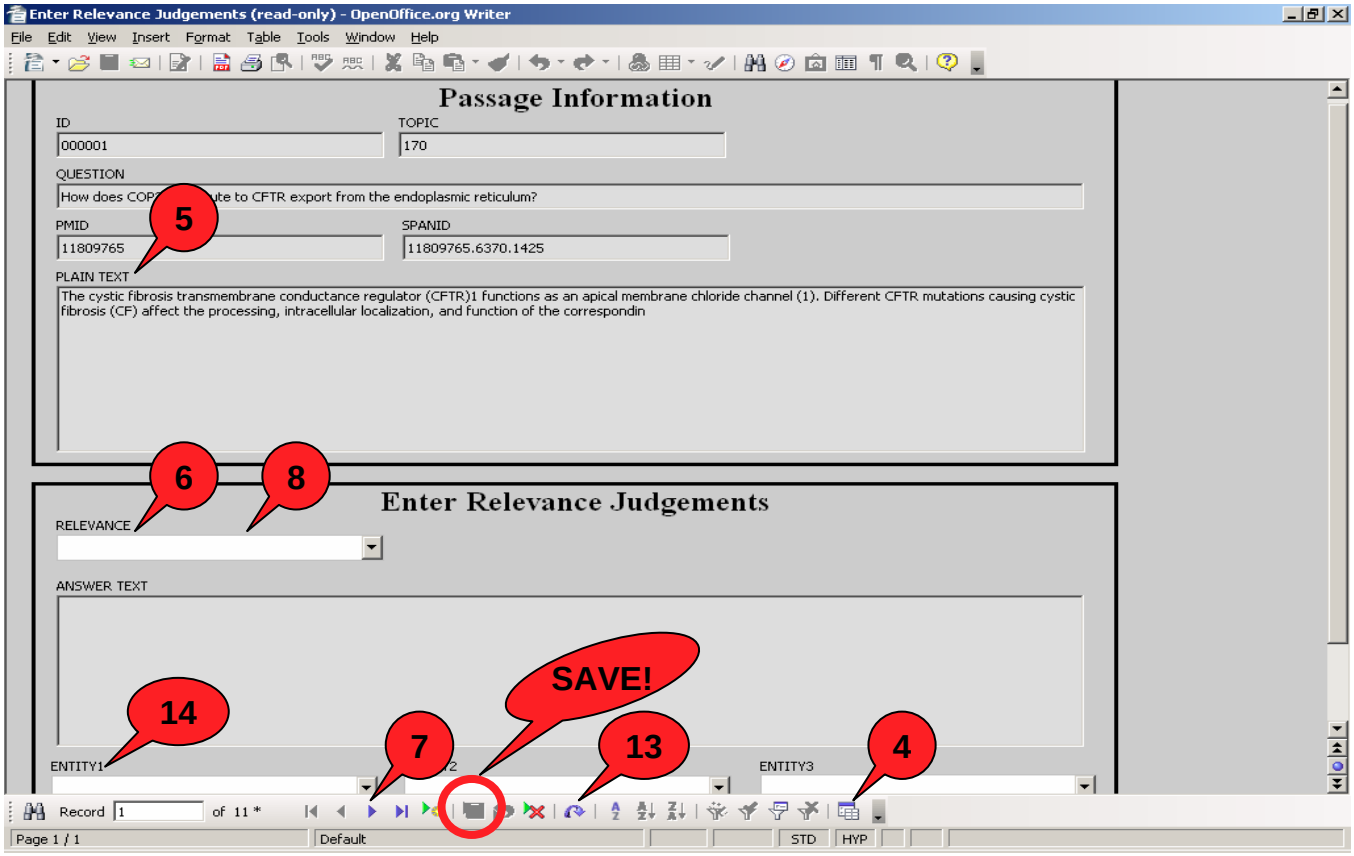
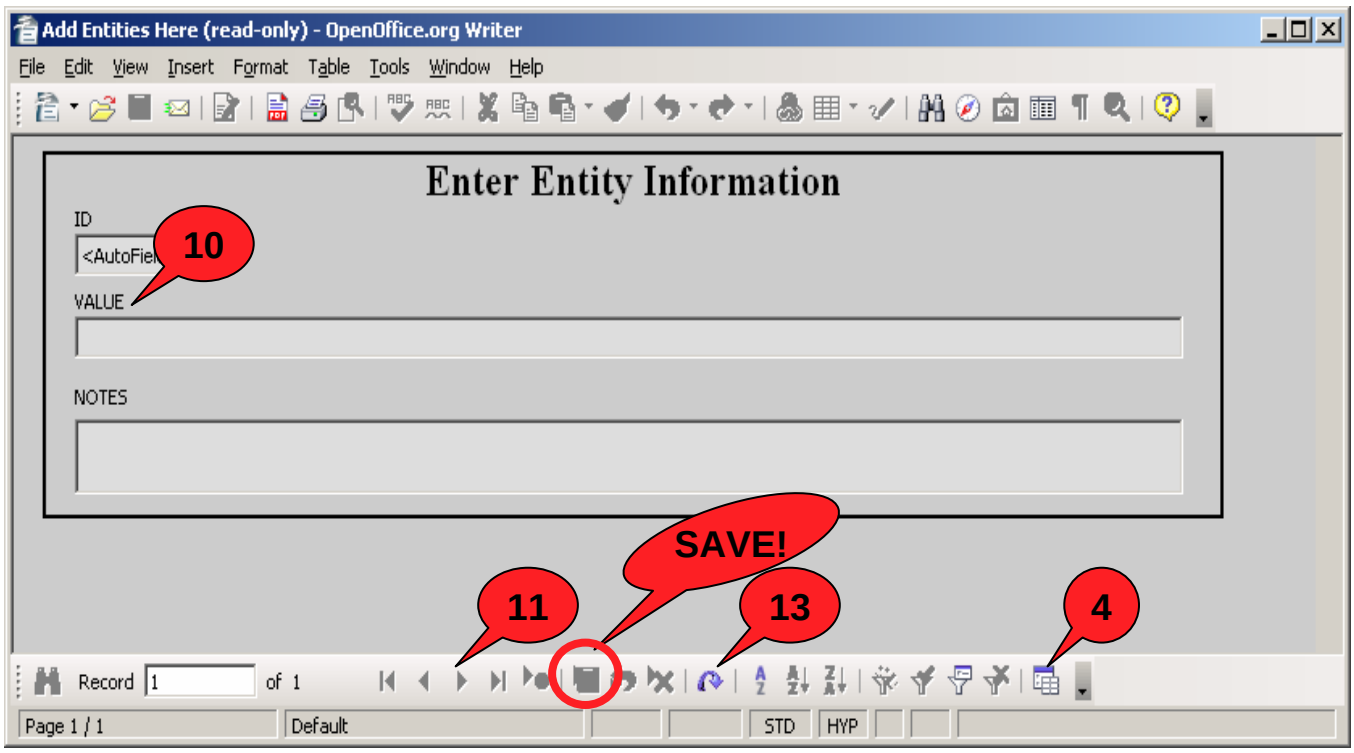


Figure 3. The Add Entities Here form (numbers correspond to steps in Part II).



Part III: Who we are and who to contact when

Phoebe Roberts, PhD. (phoebe.m.roberts@gmail.com). Phoebe oversees the judging process, having served as a judge in 2004 and 2005. She volunteers for TREC, and in her day job, she is a Text Mining Scientist with Pfizer's Systems Biology group. Contact Phoebe with scientific questions, and questions about relevance and entities. She can also help answer questions about using the Open Office Judging Database.

Aaron Cohen, MD, MS (cohenaa@ohsu.edu; http://medir.ohsu.edu/~cohenaa/). Aaron is an Assistant Professor in the Dept. of Medical Informatics and Clinical Epidemiology at Oregon Health Sciences University. He designed the Open Office database you are using to enter your judgments and he prepared the data you are judging. Aaron distributes new topics to you and he collects completed topics.

William Hersh, MD (hersh@ohsu.edu; http://billhersh.info/). Bill runs the TREC Genomics Track, now in its fifth and final year. He will see that you get paid.

Appendix I. Resources for searching (in addition to Wikipedia and Google).

MeSH – Medical Subject Headings
http://www.nlm.nih.gov/mesh/MBrowser.html
For biological processes and diseases, provides synonyms or constituent processes that are part of the indicated concept.

Medline Plus Drug Information
http://www.nlm.nih.gov/medlineplus/druginformation.html
Includes drug names and brand names, and toxicities resulting from drug administration.

IHoP – Information Hyperlinked over Proteins
http://www.ihop-net.org/UniPub/iHOP/
This database lists synonyms for proteins and provides excerpts from the literature, allowing you to familiarize yourself with the biology of the protein.

Bioinformatic Harvester
http://harvester.fzk.de/harvester/
Information from a dozen gene-centric databases (Entrez Gene, iHop, Swiss Prot, etc.) is assembled in one place, organized by gene.

AmiGO, the Gene Ontology browser
http://www.godatabase.org/cgi-bin/amigo/go.cgi
Good for brief definitions of biological processes. No disease information (see MeSH or KEGG).

BioCarta
http://www.biocarta.com/genes/allpathways.asp
List of pathways and their constituent genes, especially signaling pathways.

KEGG
http://www.kegg.com/kegg/pathway.html
Another source of pathway names and descriptions.