Questions related to the CHEMDNER task (version July 30th 2013)

What data and information will the participants receive?
We will provide a (1) small sample set with annotations and example predictions, (2) a training set consisting of abstracts and the corresponding annotations, (3) a development set consisting of abstracts and the corresponding annotations and (4) a test set consisting of abstracts.

When will a team be allowed to enter the test phase?
Participants can enter the test phase at any moment as long as it is before the submission due date of the test set predictions. You have to make sure that the submission format is correct, the BioCreative evaluation script (and documentation) and CHEMDNER sample sets help you to comply with the required format.

When will be the deadline to register as a participant team?
Participants can register as a team at any moment as long as it is before the submission due date of the test set predictions.

What will the available input formats/files be?
These files contain plain-text, UTF8-encoded PubMed abstracts in a tab-separated format with the following three columns:

1- Article identifier (PMID, PubMed identifier)
2- Title of the article
3- Abstract of the article

How will the test phase proceed?
We will place the test set articles online for download, announcing the location to participating teams (contact e-mail) and also to the BioCreative participant mailing list. Participants can then download the dataset, run their analysis and upload their annotations to a location that will be announced by the organizers before a specified deadline. Teams will have 5 days to generate five different annotations ("runs") for the test set and to submit the annotations to the organizers. You will be also asked to send a short systems description (max 2 pages) together with the results. This description should explain the used system and highlight differences between the five runs.

How sure is it that a paper talks about said chemical structure, irrespective if the specific name is right?
We do not ask the participants to return the actual chemical structures or to predict if the (main) paper (body - as opposed to the abstract) will discuss this chemical or even contain its structure. The participants only have to return the mentions of chemicals found in the article data set (i.e., title and abstract) according to annotation guidelines.

Can I adapt, retrain or integrate existing software for the recognition of chemical entities?
Yes, you can but you will have to specify what you used in the system description that we will request from the participants in order to obtain the test set.

What methods are allowed to participate?
Any method could be used, from dictionaries, machine learning, rules, regular expressions, etc.. There is no restriction as long as you do not make any manual adjustment for the predictions that you submit for the test set.

Will there be a workshop proceedings paper on my system?
Yes, we will prepare a proper volume for the CHEMDER task for the BioCreative workshop. The paper will be 2-4 in length and should be a very technical description of the your approach. You may include results obtained on the training or development set.

How were the used abstracts selected?
A detailed description on the data selection is provided in the annotation guidelines. The selection was based on subject categories from the ISI Web of Knowledge relevant for chemistry and related disciplines.

I am a little bit confused about the CHEMDNER task. It seems to me that the second task is a first step followed by the first task. Am I correct?
As for the posed tasks, there is actually no real ordering imposed and participants can submit results for any of these two or (even better) both. We have chosen to have these two separate tasks because the first one relates to associations at the document level, while the second is really a more fine-grained named entity recognition task where character offsets have to be provided. Note that the chemical document indexing task could in principle be addressed also in other ways that are different to a classical NER e.g. by using text similarity methods, text categorization, etc.. This is the reason why we did not impose a proper dependence of both tasks.

Do we have a standardized ontology to map mentions to?
No. You can nonetheless make use of any existing resource as part of your system, like the ChEBI ontology or the Jochem compound lexicon.

Should endophoric references be tagged?
No. As per the guidelines, co-reference resolution is not part of this task.

What was the background of the curators that prepared the dataset?
Curators that prepared the datasets were organic chemistry post-graduates. The average experience of the team of annotators was about 3-4 years in annotation of chemical names and chemical structures.

How was the training, development and test set selected?
Splitting into these three dataset was done by randomizing the entire dataset and then dividing it into the following collections of: 3500 (training), 3500 (development) and 3000 (test) abstracts.

Why did you not annotate proteins?
This goes beyond the scope of the task. The Gene Mention task of earlier Biocreative challenges was devoted to this entity.

In case of the chemical document indexing sub-task it is specified that: "Given a set of documents, return for each of them a ranked list of chemical entities described within each of these documents." How are these chemicals supposed to be specified?
We will ask for the same kind of output format as for the interactor normalization task (INT) of BioCreative II.5 (see the evaluation library for more details on the format).
You have to return the document identifier from which you extracted the chemical entity mention, the exact (UTF-8 encoded) character string corresponding to the chemical entity, the entity's rank on that document, and a confidence score. An example for a single document is provided below:

6780324	LHRH	1	0.9
6780324	FSH	2	0.857142857143
6780324	3H2O	3	0.75
6780324	(Bu)2cAMP	4	0.75

For the document indexing sub-task, how are the returned compounds supposed to be ranked? Based on how sure we are that a certain name refers to a chemical?
Yes. There is no order in the gold standard. Your evaluation result score will be equal no matter how you order the *gold standard* annotations. The rank (and confidence) you need to report only should express how confident you are that the extracted mention is a correct chemical entity mention according to the gold standard (i.e. annotation guidelines). If you want to show that several classifications are of equal importance, report them all with the same confidence score. However, as only the ranking influences the results, if you rank results with equal confidence differently, you might get a different score. The only factors influencing the score are how many true positives (and hence, false negatives) you have and how many false positives you do not have at each *rank*.

For the document indexing task, is the rank indicative of how important the entity is or does it represent the program's confidence that the mention is genuine?
The rank is NOT indicative of how important the entity is. It is the team program's confidence that it the returned CEM is correct.

Does the confidence score we have to report express how likely it is that the specific

name refers to a structure one has mapped the name to?
No. We do not request this mapping and the confidence score (and rank) only express how likely it is that a tagged name truly is a chemical entity mention according to the annotation guidelines.

For the document indexing task I understand that you essentially require the entities found by the chemical entity recognition task in ranked order, uniquified to those mentions with different Strings.
You are right. For the document indexing task you have to provide the list of *unique* chemical entities, that is each different entity mention in the article, as defined at the String level (just as if using the UNIX command “uniq”). These unique entities then have to be ranked by your system according to your confidence in the extracted String representing a chemical entity. (We do not ask for any mention offsets in this task, while we do not ask for any ranks in the entity recognition task.)

Then, what is the confidence score for?
It is not really relevant for your individual results/scores. Your score will be established using the ranking you provide on your results. But we might use this score to evaluate the performance of a combined meta-annotation created from all participants and for similar analyses of the results.

For the document indexing task, the rank should be based on the confidence?
Yes, the rank in principle should be based on confidence your system has in the result. I.e., the higher the confidence (in the (0,1] range) of a result, the higher its rank should be.

For the document indexing task, only the entity names should be extracted?
Yes, thats right. You have to return a non-redundant ranked list of chemical entities. You should NOT provide the exactly same mention multiple times.

For the document indexing task, can I provide multiple ranks for a chemical name?
No. Chemical entities for the CDI task have to be unique. You can test this assertion by using the evaluation script on plain-text formatted results - it will abort with an error if your results are not unique.

How many chemical entities for a given abstract can I return for the document indexing task?
We advise you not to return to long lists of entities, keeping in mind the practical importance of the underlying systems. We advise you to return less than 100 unique chemical entity mentions for an abstract. Longer lists with many irrelevant results will penalize the precision score of your system.

How many runs can I submit for the document indexing or the entity recognition subtask?
You are allowed to send up to 5 runs for each task, that is 5 for the CDI and 5 for the CEM.

Do I need to submit results for both subtasks?
You do not need to provide submissions for both, although we encourage to try to return predictions for the CDI and the CEM.

For the CDI should the returned compounds be mapped to actual chemical entities, i.e. structures?
Not this time. We want to carry out such a normalization evaluation in the future. As existing database do not cover all the chemical entities that are mentioned in the literature or patents, we simplified the task for this first time.

For the CDI subtask do we need to return the compounds 'normalized', which identifiers should be used? InChi? SMILES? PubChem? CHEBI?
Not this time.

For the CDI task are the compounds derived from the abstracts or the full text articles?
The annotations are restricted to the abstracts only.

What evaluation measures will be used for the CDI task?
For the CDI, in addition to precision, recall and particularly the balanced F-score, we will measure the Average Precision (AP) to evaluate the ranking performance. For more information, please refer to the evaluation library documentation web. You should probably optimize your system against the combined, "F-measured Average Precision" score to do well on both scores (F-measure and AP). (http://www.biocreative.org/resources/biocreative-ii5/evaluation-library)

What evaluation measures will be used for the CEM task?
For the CEM, we will use precision, recall and mainly the balanced F-score.

For the CDI task, does one have to first recognize all the chemical mentions with start and end indices, normalized them into entities and then rank them based on this confidence score?
For the CDI we will not ask you to normalize the entities and also there will be no offsets requested.

So the CDI task is the one to match each document onto a given lexicon (entities). Are the gold standard entities ranked on article or set level? Is there descriptive definition for each lexicon entry?
The CDI task consists in providing - for each document - a ranked list of chemical compound names, rather than matching a document to a lexicon. There is no "official" lexicon of "allowed" mentions for this task (other than the annotation guidelines). You are of course allowed to use any lexical resource you want, for instance PubChem, ChEBI, Jochem, Chemspider, etc.. The entities provided in the gold standard are not ranked in any way, they are simply the set of all

valid (and unique) chemical entity mentions found in the article according to the annotation guidelines.

The CEM sub-task specified that: "Provide for a given document the start and end indices corresponding to all the chemical entities mentioned in this document". It would seem that this second task is a prerequisite for doing the first task. But unless the first task does involve mapping the names to structures, it is not clear what the difference is between the two tasks.
The difference between the two tasks is that in case of the CDI, you have to return *unique* and *ranked* mentions, but we do not ask you for their offsets. On the other hand, for the CEM task, we do not ask for a rank, but *all* mentions (i.e., not just the unique mentions) with their *offsets* should be reported.

I think there might be a possible error in one of the sample annotations: "trisaccharide" is annotated, rather than the full "aminoglycoside trisaccharide".
No, this is in line with our guidelines (see page 5): aminoglycoside is a general term (N3), while trisaccharide is an allowed CEM. Therefore, this is an annotation in accordance to the guidelines.

In the CHEMDER annotation guidelines, point M1: "substituted" in "N-substituted-2-alkyl-3-hydroxy-4(1H)-pyridinones" shouldn't be annotated (personally I think "substituted" should have been annotated in all cases, as it gives the structural information that a hydrogen has been replaced by a substituent)
According to the annotation guidelines, "substituted" should be annotated if inside a chemical entity, but not if it is provided as an isolated word. In the example above, the word "substituted" is inside the chemical term (as opposed to, for example, "substituted quinolines"), and therefore must be annotated.

In the CHEMDER annotation guidelines, the annotation of "Pd/C" as "Pd" "C" may be inconsistent with O6 if it is considered as an abbreviation of palladium on carbon. "Pd/C" describes a single concept and so should be annotated as one entity in my opinion (although for consistency one would then need to tag "palladium on carbon" [TRIVIAL] as one entity)
They are not abbreviations, we consider them two chemical formulae. As for the question if this is one or two entities... this is a rather tricky point. The questioner understands that this is a single chemical, as it is a mixture, but we decided to keep them as two separate entities in our curation rules. The other alternative, differing from the "part of CEM that is part of …" would require very conscious annotations (that in turn would be very sensitive to errors, too).

In the CHEMDER annotation guidelines, there are some idiosyncrasies (probably not mistakes, but a bit odd). Small cyclic peptides are allowed, but no cyclic carbohydrates.

We agree, but we thought a threshold is required here.

In the CHEMDER annotation guidelines, biochemical polymers are not annotated, but synthetic polymers are.
Accepting all biochemical polymers means accepting all proteins - and this is totally outside of the scope of this task, as this entity type has its "own" biological world. On the other hand, discarding synthetic polymers was odd because chemists traditionally regard them as chemical entities. Moreover, they are not as frequent as proteins and glucosides.

In the CHEMDER annotation guidelines, aqueous as an adjective is ignored but included as a state symbol e.g. CuSO4(aq) This is really a difficult issue for us and we have discussed this topic several times.
Being strict, you might be right, but annotating it as adjective would add information that we do not want to capture from a pragmatic point of view. In the current version we will remain with this convention.

The annotation guidelines mention subcategories of family e.g. FAMILY->SYSTEMATIC and FAMILY->TRIVIAL but these aren't present in chemdner_sample_annotations.txt. Were the subcategories just there to help illustrate the different types of entities that fall under FAMILY?
You are right. We plan to carry out such an annotation only for a subset of the (test set) data, due to the additional workload. If possible, we might then extend this effort to the entire set. We can not provide any dates for this effort at the moment.

In the CHEMDER annotation guidelines, point N7 it is stated that lead is excluded from the annotations. Does that extend to its chemical use, e.g. "lead acetate"?
Yes. The benefit of removing this tricky term compensates for the times we might miss the chemical term. Therefore, this decision should make it easier for participants, as they only need to define a corresponding “stop-word” list.

In the CHEMDER annotation guidelines, Something like "5-(1-azidovinyl)uracil derivatives" technically describes a FAMILY. My reading of CEM-1 and the sample annotations is that SYSTEMATIC is preferred with the word derivatives being ignored. Is this correct?
Yes, you are right. This was done to make the task easier (less context-dependent) for machine learning.

In the CHEMDER annotation guidelines, CEM-3 (other common names) and the sample annotations annotate chemical names appearing in the names of enzymes. These occurrences don't necessarily indicate anything about the structure of the enzyme and instead relate to the function of the enzyme e.g. "pyruvate kinase". Assuming this was intentional, was this done to make the task easier for the machines, human annotators or

both? (the annotation of thiophene in "Gewald thiophene synthesis" is a similar issue, as this refers to a procedure rather than a chemical; yet another is 1H in the context of proton NMR)
Yes, to make it easier for the machines (and also, in some cases, for the human annotators).

Will there be a special issue published in a journal related to the CHEMDNER task?
Yes, we plan to make a special issue on the CHEMDER task, in a similar way as had been done for earlier BioCreative challenges.

Is it possible that a single person can be part of several participating teams?
In principle, there is no particular restriction in terms of the number or association of members per team. However, the BioCreative website only allows one user to be associated to one team and it is that team she will be associated with in all official communications.

I have found some missing annotation in the dataset, what should I do?
Please contact the organizers and provide them with the details of the missing annotation.

I am a chemist, is there some possibility to help in preparing the annotations?
Yes of course, in case you would like to help, please contact the task organizers. The annotation interface we constructed is easy to use. The most important issue is to understand the annotation guidelines properly.

Who should I contact in case of doubts about the task?
Please send your questions via e-mail to the task organizers or consider using the BioCreative participant mailing list in case it is an issue of common interest.

Does the CHEMDER task have a closed setting in the sense of only being allowed to use the provided training collections?
No, you are allowed to use any existing resource to augment your predictions, with the exception of doing manual annotations.

Can I use both the training and development set to train/tune/implement my system or must I use only the training set?
Of course you can use both, we have provided two separate datasets because of timing issues.

Are the evaluation results anonymous?
The obtained results are not anonymous. If you download the CHEMDNER test set we expect that you agree to make the obtained results and team public. This is not a competition but a community challenge with the aim of learning together how to improve text mining systems and sharing at least very basic information on the used systems. If you have a problem with making your data public, please contact the organizers before downloading the test set.

Is there some library or script I could use to evaluate my system?
Yes, we provide the BioCreative evaluation library available in the resource section of the BioCreative webpage.

It would be interesting to know how good the inter-annotator agreement is, as that defines an upper bound of how well an automated solution will perform.
That is right, we will examine the outcome of multiple annotators and the corresponding IAA. We are of course aware that the IAA at the level of the individual CEM classes will be considerably lower than that of the mentions themselves. These numbers should be provided at the evaluation workshop.

What will the survey that forms part of the result submission be asking from the teams?
A short, technical description of your approach (e.g., machine learning and NLP methods, used resources and lexica, etc.).