Project ChEMU

Guidelines for Annotating Chemical Entities and Events related to Chemical Reactions

Version <1.02>

Revision History

Version Number, Date	Revision made by	Change Summary
1.00	Christian Druckenbrodt	created
1.01	Christian Druckenbrodt	adjusted
1.02	Christian Druckenbrodt	adjusted

Contents

SCOPE OF THE GUIDELINE

This document is the manual for the Full-Text Patent Annotation (FTPA) project and should advise how to annotate manually (or check pre-annotated) parts of a chemical reaction, e.g. chemical entities, reaction conditions etc. in patents correctly. In addition, events connecting the entities must be identified and annotated. It will provide definitions, and examples, in which cases and how a string of characters embedded in text of patents must be annotated accordingly. The resulting manually annotated BRAT files can be used as a so-called “gold standard” in order to determine Recall and Precision values regarding corresponding automatic annotations, produced by state-of-the-art text mining tools.

OVERVIEW

The recognition of chemical reactions is an essential step in the extraction of chemical information from any kind of document. This goes beyond the annotation of chemical entities as it also covers their role within the context of a chemical reaction. In addition, some typical conditions should be annotated which provide contextual hints to the presence of a chemical reaction, whilst also providing additional information beyond a simple starting material/product relationship. Finally, some of these entities should be connected by annotating events. In the current FTPA project we only focus on representations of entities, conditions, and events as strings in snippets of running text. Drawn chemical entities/structures, conditions, and events in pictures, schemes or tables are excluded, as state-of-the-art technologies are currently not able to reach enough quality in such regions of documents.

Determination of acceptability for chemical entity strings in a reaction derived from snippets of running text is often quite difficult. The complexity originates from the matter of fact, that chemical entities present in a reaction snippet may or not may not play a certain role in a reaction. The problem here is to identify all chemical entities playing a role in the reaction in the first place followed by the classification of these chemical entities according to the role they have in the chemical reaction.

An additional problematic issue is that strings used in snippets often refer to context beyond the individual snippet. The string used in the snippet may only be a representative of the complete chemical entity. Strings of chemical entities themselves can be detected, annotated, and to some extend resolved by state-of-the-art name-to-structure tools. Representatives, be it a label or a reference to the “product of a certain reaction”, can be detected and annotated within this context but they cannot be resolved without any coreference.

As the basically only snippets of reaction texts will be part of the to be annotated text one further issue is not so prominent but is still present: in context/within natural language may change their intrinsic

meaning depending on the different environments. That means, from an atomistic point of view, a string/word/term/group of characters may have an intrinsic unique meaning in the chemical domain (e.g. “Glucose”; “gold”), but embedded in running text it might mean something different (e.g. “Glucose transporters are a wide group of membrane proteins”) although it is still the same domain or even in other domain (“The exchange rate under the gold standard monetary system…”). As a consequence, documents, frequently annotated automatically by machines, will contain a huge amount of strings annotated as chemical entities. Although all of these annotations are somehow related to chemistry, only a certain amount of them are actually contextual chemical entities. Due to the above explained contextual constraints, the embedded meaning for the rest of the annotated strings in running text is not a chemical entity but something else.

Events are here defined as descriptions of actions that combine related entities, like starting material A is added to starting material B. Here “starting material A” and “starting material B” are chemical entities combined with the event “added” In general events are represented by verbs, however if better fitting nouns can be annotated, too.

INTRODUCTION TO PATENTS

A patent is a right granted to the owner of an invention that prevents others from making, using, importing or selling the invention without his permission. A patentable invention can be a product or a process that gives a new technical solution to a problem. It can also be a new method of doing things, the composition of a new product, or a technical improvement on how certain objects work. For an invention to be patentable, it must, in general, satisfy three key criteria: New, inventive step, and industrial application.

The sections of patents are quite conserved: title, bibliographic information (patent number, dates, inventors, assignees, IPC classes, …), abstract, description, and claims. Most of the chemical data will be found in the experimental section of the description, while compounds claimed (protected) are available in the claims section. Drawings, sequences, or other additional information will normally be found at the very end of a patent.

This particular project concentrates exclusively on the experimental section in the description as the source of reaction text snippets.

HIGH LEVEL RULES

In general, chemical reaction is a process leading to the transformation of one set of chemical substances to another. Here, the task is to identify specific types of chemical compounds, i.e. to assign the label of a chemical compound and its connecting event according to the role the chemical

compound plays within a chemical reaction. Mostly the reaction full text can be divided into actual reaction and the following work-up – in the reaction the product is completely formed. In the work-up the product remains unchanged and steps undergone are for isolation of the product only. The actual reaction and the following workup must be kept separate in the annotation and are this defined by two different kinds of events. The presence of a chemical entity in one of the separate parts has consequences for the annotation.

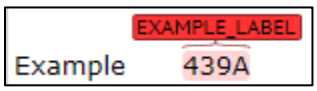
All snippets represent reaction full texts and are already pre-tagged. All tags must be revised and – if necessary – corrected. Missing tags must be set. Entities present more than once – also in different representations – must be annotated at each occurrence.

Part 1: Entity Annotation

We define 10 different entity types, including:

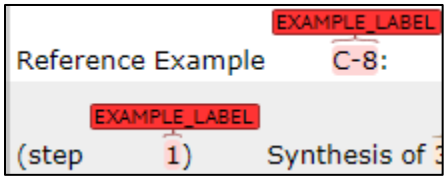
EXAMPLE_LABEL, REACTION_PRODUCT, STARTING_MATERIAL, REAGENT_CATALYST, SOLVENT, TIME, TEMPERATURE, YIELD_PERCENT, YIELD_OTHER and OTHER_COMPOUND.

EXAMPLE_LABEL



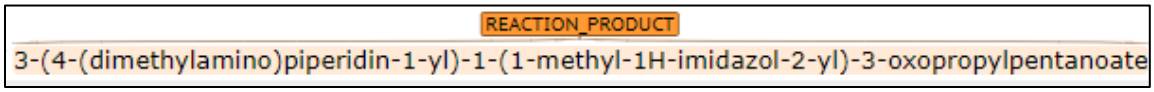
If a reaction label is given in the snippet this must be annotated with this tag.

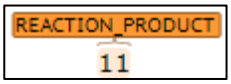
Words like “Example”, “Step”, “Intermediate”, “Core” and “Reference example” must not be annotated, neither are parentheses, brackets or braces. Do not confuse reaction labels and compound labels. This tag must only be annotated to reaction labels.



There can be more than one reaction label in a snippet as reaction labels consider both the label of the finally generated product, as well as, the described intermediate product.

REACTION_PRODUCT

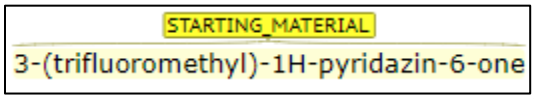


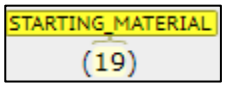
, or


A product is a substance that is formed during a chemical reaction and must be annotated with this tag.

As REACTION_PRODUCT all representations of the product must be annotated, product name and/or label but also representatives like “title compound” generally pointing to the title of the reaction snippet. All such variants of the reaction product must be annotated in the snippet.

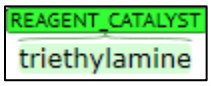
STARTING_MATERIAL

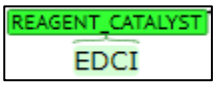


, or


A substance that is consumed in the course of an organic chemical reaction providing carbon atoms to products is considered as starting material and must be annotated with this tag. In inorganic reactions the prerequisite is that the starting material provides any atoms to the product.

REAGENT_CATALYST

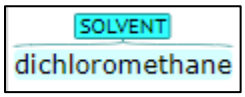


, or


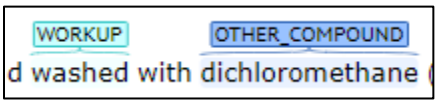
A reagent is a compound added to a system to cause or help with a chemical reaction. Compounds like catalysts, bases to remove protons or acids to add protons must be annotated with this tag. In organic reactions compounds providing non carbon atoms to a product are also considered REAGENT_CATALYST.

Reagents must be involved in the reaction. Compounds given in the course of work-up, like MgSO4 for drying must not be annotated with this tag. These must be annotated as OTHER_COMPOUND. It may well be that within the same snippet compounds may have a role as REAGENT_CATALYST and OTHER_COMPOUND.

SOLVENT



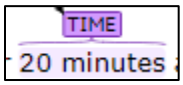
A solvent is a chemical entity that dissolves a solute resulting in a solution. The solvents used in the reaction must be annotated with this tag. In case of solvent mixtures all individual solvents must be annotated. Solvents that also have the role as STARTING_MATERIAL must only be annotated as such. Solvents used in the work-up must not be annotated as SOLVENT but as OTHER_COMPOUND.



It may well be that within the same snippet compounds may have a role as SOLVENT and OTHER_COMPOUND.

Alternatively used solvents must not be annotated (e.g. in case of ‘dichloromethane or chloroform was used’ none of the mentioned solvents must be annotated).

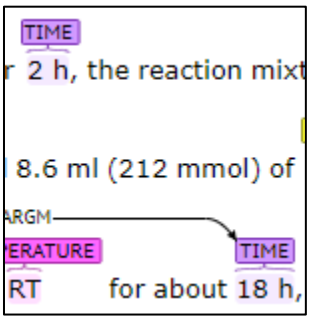
TIME



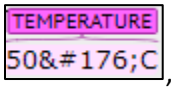
The reaction time of the reaction must be annotated with this tag.

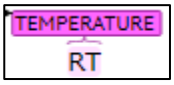
If just one particular time information is given (e.g. 20 min, 2 h or 3 d), that time information must be annotated.

If different procedures with different reaction times were carried out consecutively (e.g. 30 min stirring at 20 °C, then 2 h reflux) the individual times must be annotated.



TEMPERATURE



, or


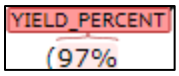
The temperature (range) at which the reaction was carried out must be annotated with this tag.

In case a reaction was carried out at more than one temperature, the given lowest and the given highest temperature must be annotated. Temperatures within this range must not be annotated.

The temperature range can be given with (1) numerical values or (2) by a specific keyword. “room temperature (RT)”. This keyword must be annotated if the temperature is not specified by a concrete value. “Room temperature” is representing 20 °C here. This must be considered when setting ranges and entering only the minimum and maximum temperature.

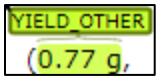
It must be kept in minds that the xml expression for °C is given as “°”.

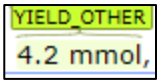
YIELD_PERCENT



Yield given in percent values must be annotated with this tag. Only isolated yields must be annotated with this particular tag.

YIELD_OTHER



, or


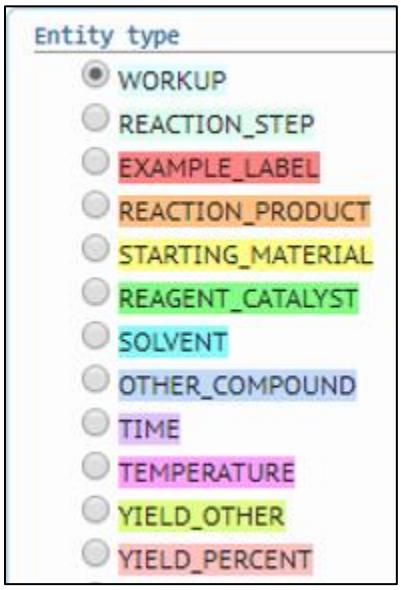
Yields provided in other units than % must be annotated with this tag. YIELD_OTHER can be given as amount of substance (in mol or mmol) received mass of substance (in g or mg)

OTHER_COMPOUND

OTHER_COMPOUND silica gel, or OTHER_COMPOUND [4-[6-(trifluoromethyl)pyridazin-3-yl]-2-pyridyl]methylamine hydrochloride

Other chemical compounds, which are not the products, starting materials, reagents and solvents, must be annotated with this tag. This can be standard chemicals used for drying but also the overall title compound of a set of reactions present in the snippet.

The following colors are used to distinguish different entity types in BRAT.



Real Live Problematic Examples

All the above entities are generally pre-tagged either by correlation with Reaxys content or by a chemical entity recognition process. Some mistakes are intrinsically present in the underlying processes. Therefore, the main aspect of the chemical entity annotation will be a thorough check of existing tags. Greatest care must be taken to ensure consistency and correctness.

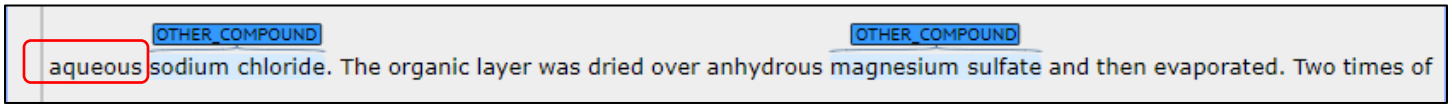
In general, three aspects have to be considered here:

- Missing tags – that must be annotated
- Wrong tags – that must be changed
- Wrong tags - that must be deleted

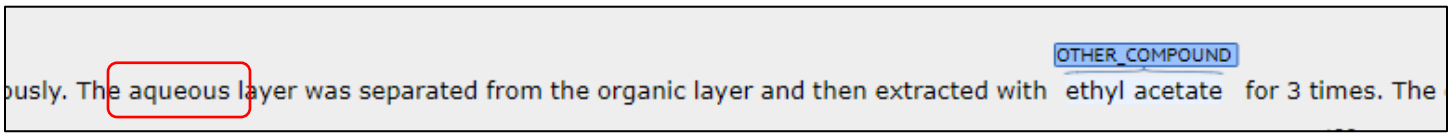
Missing tags

4 To a solution of STARTING_MATERIAL 5-nitrosalicylic acid (1.00 g, 5.46 mmol) dissolved in 15 mL of DMF, REAGENT_CATALYST K2CO3 (4.52 g, 32.8 mmol) and STARTING_MATERIAL iodoethane (2.2 mL, 27.5 mol) were added, and the mixture was stirred at TEMPERATURE 50°C TIME for 6 hours. The solvent was

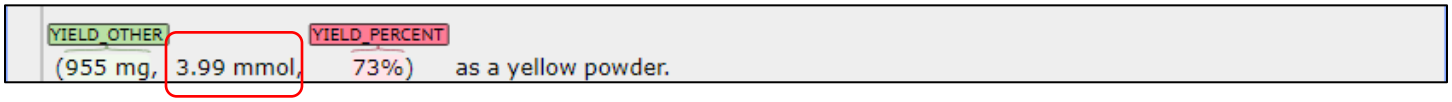
The solvent DMF – dimethylformamide is missing in the original annotation. The mapping process, as well as, the chemical entity recognition are mostly depending on systematic names. Abbreviations like DMF here may cause problems and remain un-tagged.



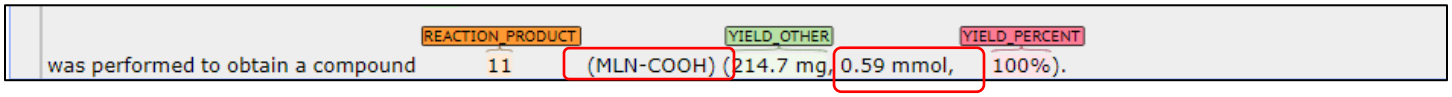
The solvent water is “hidden” in this specific term. It is not detected by chemical entity recognition.



The term “aqueous” can be ambiguous. Only in cases where this word is used as a placeholder for the compound annotation must be done. In the above case the “aqueous” is only further identifying the layer and must therefore not be annotated.

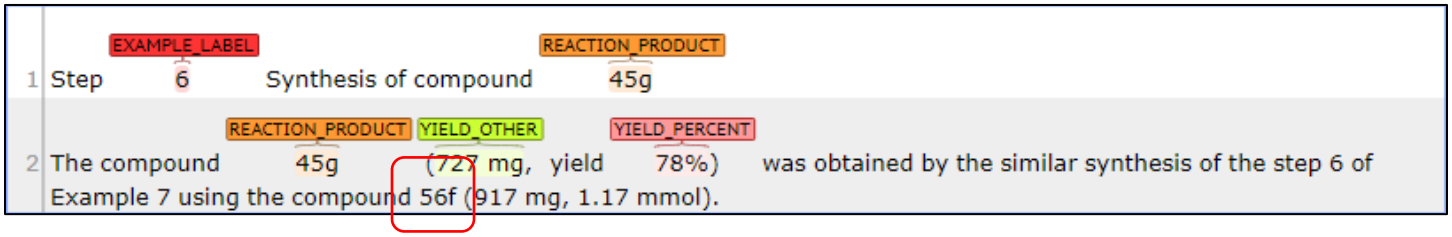


The absolute yield provided in mmol is not pre-annotated. Please add further annotation YIELD_OTHER.



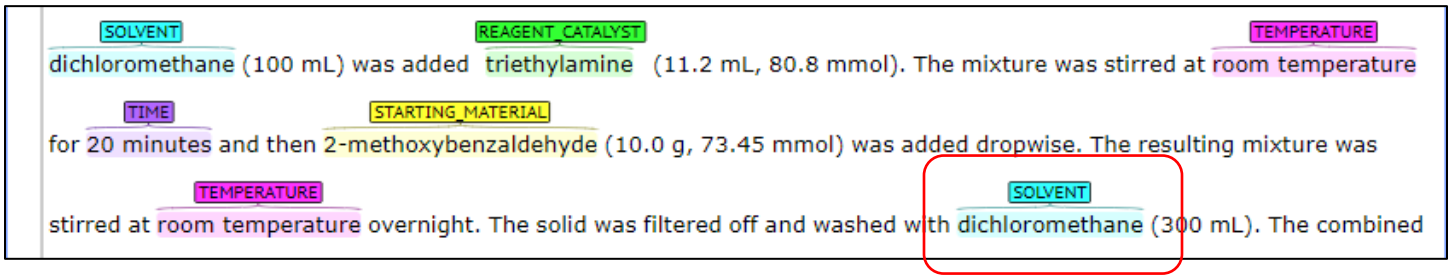
The second label provided for the product is missing and must be annotated. In addition, the absolute yield in mmol is missing.

In case the reaction is missing the starting material and shows only the REACTION_PRODUCT as indirect entry the STARTING_MATERIAL should be identified semantically.

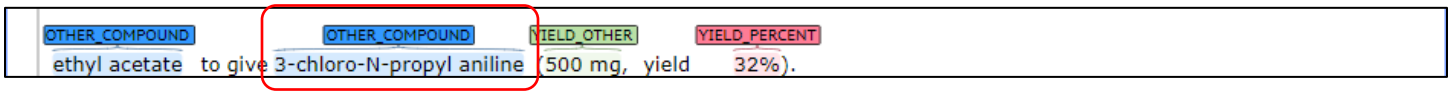


Here no STARTING_MATERIAL was pre-tagged. By semantic analysis of the snippet 56f can be identified as STARTING_MATERIAL.

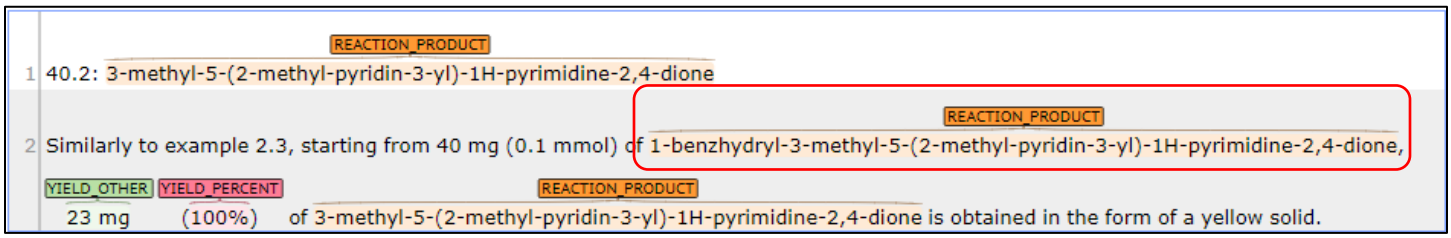
Wrong tags - that must be changed



Dichloromethane is present in two instances. Whilst the first instance is correctly tagged as SOLVENT the second instance is clearly part of the workup (washed with…) and must therefore be annotated as OTHER_COMPOUND.



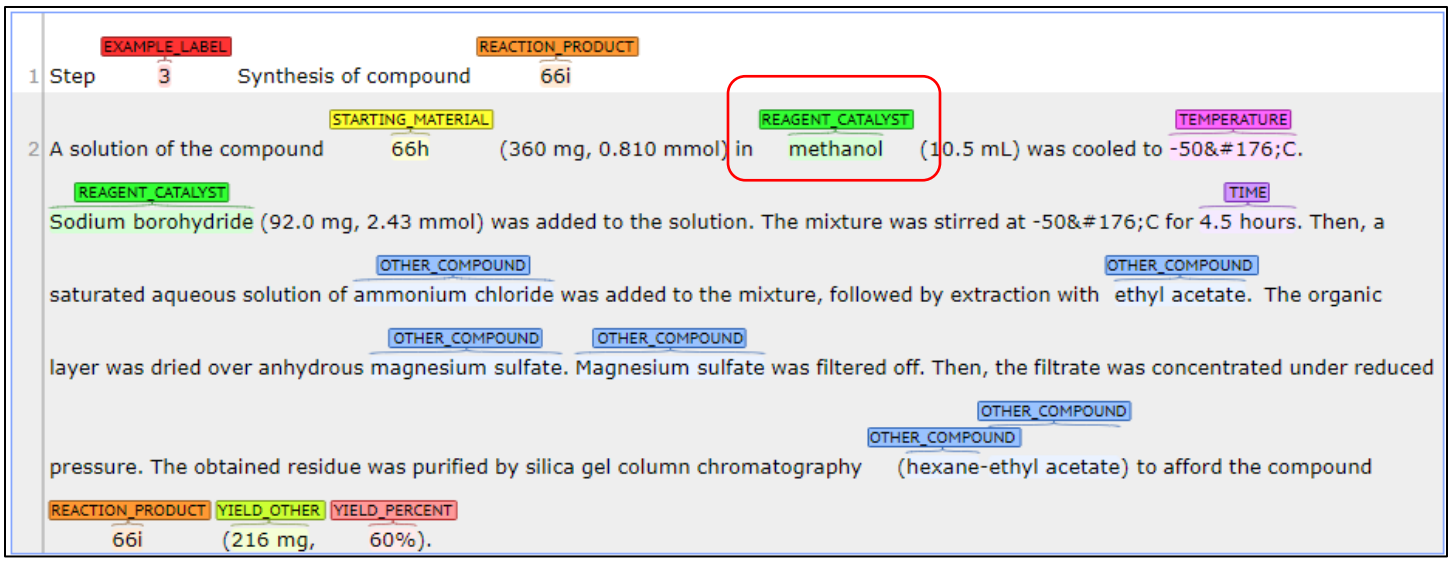
3-chloro-N-propyl aniline has been tagged as OTHER_COMPOUND but must have been annotated as REACTION_PRODUCT



1-benzhydryl-3-methyl-5-(2-methyl-pyridin-3-yl)-1H-pyrimidine-2,4-dione has been tagged as REACTION_PRODUCT but must have been annotated as STARTING_MATERIAL

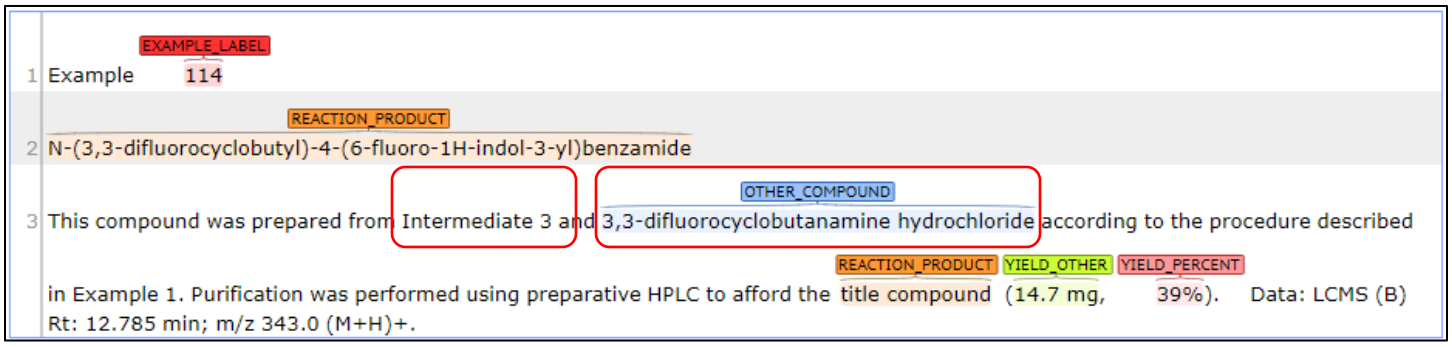
Notes:

- In case the reaction in the snippet holds annotations for main STARTING_MATERIAL and REACTION_PRODUCT which are not directly given no changes regarding assignment of the chemical entities must be made within the reaction regarding REAGENT_CATALYST, SOLVENT and further STARTING_MATERIALS. These entities are pre-tagged by match with data excerpted for Reaxys. Without any context present there practically no chance to clear up matters.



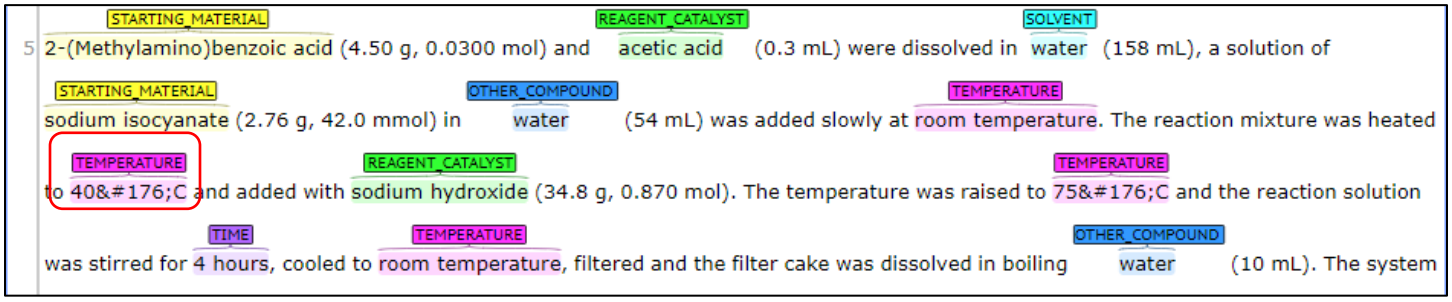
Methanol is also a typical SOLVENT but according to pre-tagging it is a REAGENT_CATALYST. No changes can be made here.

- If either STARTING_MATERIAL or REACTION_PRODUCT is given directly necessary changes must be made. These changes are often necessary when the reaction used for pre-tagging was an “analogue reaction”. As for these no starting materials are entered by default starting materials are tagged by the chemical entity recognition only as OTHER_COMPOUND.

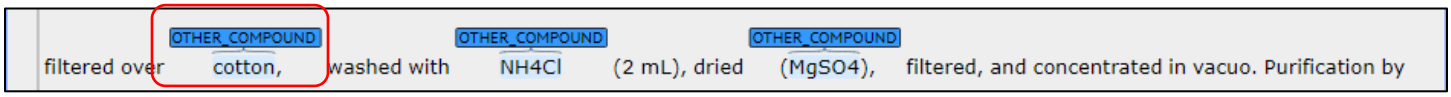


Intermediate 3 has not been detected at all and 3,3-difluorocyclobutanamine hydrochloride only as OTHER_COMPOUND. Both must have been tagged as STARTING_MATERIALS

Wrong tags - that must be deleted



40°C was tagged as TEMPERATURE. In this case a range is given with the lowest temperature at 20 °C (room temperature) the highest point with 75 °C. Therefore, the tag for 40°C must be deleted.



Cotton is no chemical compound. This wrongly assigned tag must be deleted.

Part 2: Event Annotation

OVERVIEW

A chemical reaction leading to an end product usually consists of a sequence of individual event steps. In an event extraction task those steps should be identified based on the given gold chemical entity annotations set (as defined in the first part of these guidelines). To enable this task the events must be annotated as part 2 of these guidelines.

In the event annotation task event trigger words (e.g. “added”' and “stirred”') must be annotated. Based on their function and location within the snippet, they must be labeled by either as REACTION_STEP or as WORKUP. In addition, trigger words must be linked with their related chemical entity. The annotation of the relationship between two event triggers is out of scope in this task. As trigger Words already imply, only single words must be annotated regardless of their grammatical form, e.g. auxiliary verbs must not be annotated. These trigger words are mostly verbs like “added” or “washed” but can also be nouns like “purification” if no such contextual verb like “purified” is present in the text.

REACTION_STEP

REACTION_STEP charged, or REACTION_STEP added., or REACTION_STEP stirred, but also REACTION_STEP obtain, or REACTION_STEP provide, or REACTION_STEP yielding

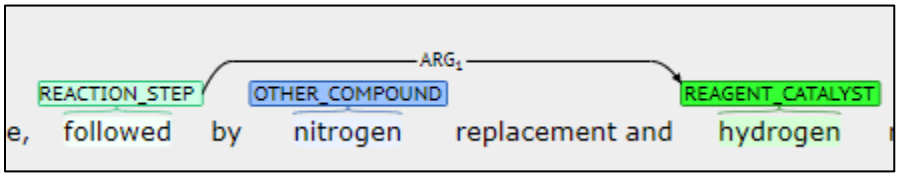
Within the REACTION_STEP event, the starting materials are converted into the product.

WORKUP

WORKUP washed, or WORKUP dried, or WORKUP purified

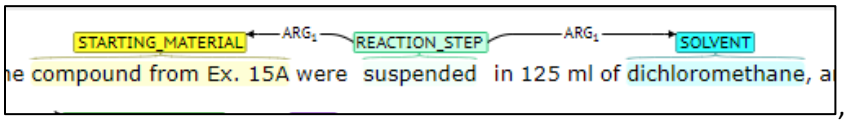
Within the WORKUP event, the chemical product is only isolated, i.e. this event type refers to the series of manipulations required to isolate and purify the product(s) of a chemical reaction.

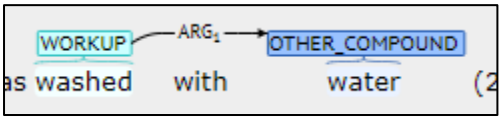
No further events apart from these two must be extracted as they do not below to the reaction context, e.g. characterization must not be extracted. Inert-gases, not taking part in a reaction must be annotated as chemical entity, here as OTHER_COMPOUND, but must not be linked to an event.



For annotation, we adapt semantic argument roles **Arg_1** and **Arg_M** to label the relations between the trigger words and the chemical entities.

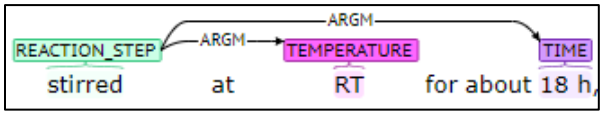
Arg_1

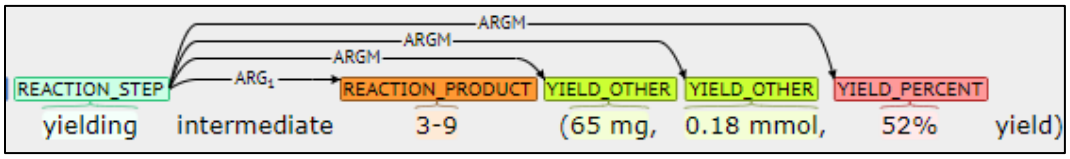


, or


Arg_1 is used to label the relation between an event trigger word and a chemical compound.

Arg_M



, or


Arg_M represents adjunct meanings, which is used to label the relation between a trigger word and a temperature, time or yield entity.

The linking must be done starting from the event drawing the arrow to the chemical entity.