

```

setwd(iframe)
files <- list.files(pattern = "pdf$")

all<-filter(tlist,!is.na(tlist$CAS))
all<-all[,c(1:2,5:ncol)]
##loop 1 - file
for(file in files){
#  cas<-word(file,6,sep=fixed(' '))
  cas<-word(file,2,sep=fixed(' '))
  cas<-word(cas,1,sep=fixed('.'))
  for(row in 1:nrow(all)){
    if(all$CAS[row]==cas){

t1<-pdf_text(file)
t2<-strsplit(t1,"\n")
t3<-t2[[1]]
for(l in 2:length(t1)){
  t3<-append(t3,t2[[l]])}

#Remove table of contents, aggr section, references
t3<-t3[t3 !='']
rows<-c(grep("Toxicological information",t3),grep("References",t3))
if(!is.na(rows[2]) & !is.na(rows[4])){
t4<-t3[rows[2]:rows[4]-1]

#Identify header sections, start and stop
head_rows<-which(t4 %in% c("Toxicokinetics, metabolism and
distribution",
                        "Acute Toxicity",
                        "Irritation / corrosion",
                        "Sensitisation",
                        "Repeated dose toxicity",
                        "Genetic toxicity",
                        "Toxicity to reproduction",
                        "Carcinogenicity",
                        "Specific investigations",
                        "Exposure related observations in humans"))
headstop<-c()
for(a in 1:(length(head_rows))) {
  if(a<length(head_rows)) {headstop[a] <- head_rows[a + 1] - 1}
  else{headstop[a]<-length(t4)}}

#loop for DNEL output
admin<-t4[1:(head_rows[1])]
ind<-grep("Hazard via",admin)
genind<-grep("General Population - Hazard via",admin)
if(length(genind) != 0){
genstop<-c()
for(c in 1:(length(genind))) {
  if(c<length(genind)) {genstop[c] <- ind[grep(genind[c],ind) + 1] - 1}
  else{genstop[c]<-length(admin)}}

for(m in 1:length(genind)){
gen<-admin[genind[m]:genstop[m]]

```

```

#Identify DNEL sections, start and stop
dnel_rows<-which(gen=="Hazard assessment conclusion")
if(length(dnel_rows) != 0){
  dnelstop<-c()
  for(c in 1:(length(dnel_rows))) {
    if(c<length(dnel_rows)) {dnelstop[c] <- dnel_rows[c + 1] - 1}
    else{dnelstop[c]<-length(gen)}}

##loop a - separate DNELs
for(c in 1:length(dnel_rows)){
  if(gen[dnel_rows[c]+1]=="DNEL (Derived No Effect Level)") {
    dnel<-gen[dnel_rows[c]:dnelstop[c]]
    acute<-0
    syst<-0
    update<-matrix(0,1,ncolt-4)
    update<-cbind(cas,dnel[1],update)

    #Identify sections and subsections, start and stop
    sect_rows<-which(dnel %in% c("Most sensitive endpoint"))

    if(length(sect_rows!=0)){
      ##loop b - sections
      for(d in 1:length(sect_rows)){
        if(dnel[sect_rows[d]+1]=="acute toxicity"){acute<-1}
        if(dnel[sect_rows[d]+1] %notin% c("acute toxicity","")){syst<-1}
      } #end loop b
    }

    if(acute==1){update[,c(54,55)]<-1}
    if(syst==1){update[,c(54,56)]<-1}

    #add to row of skeleton table
    all[row,3:(ncolt-2)]<-all[row,3:(ncolt-
2)]+as.numeric(update[1,3:(ncolt-2)])
  }
} #end if
} #end loop a
}
} #end DNEL loop

##loop 2 - headers
for(a in 1:length(head_rows)){
  if(head_rows[a] != headstop[a]){
    head<-t4[head_rows[a]:headstop[a]]

    #Identify subheader sections, start and stop
    subhead_rows<-which(head %in% c("Basic toxicokinetics","Acute
toxicity: oral",
                                "Acute toxicity: inhalation","Acute
toxicity: dermal","Acute toxicity: other routes",
                                "Skin irritation / corrosion","Eye
irritation",

```

```

        "Skin sensitisation", "Respiratory
sensitisation", "Repeated dose toxicity: oral",
        "Repeated dose toxicity:
inhalation", "Repeated dose toxicity: dermal",
        "Genetic toxicity in vitro", "Genetic
toxicity in vivo", "Dermal absorption",
        "Developmental toxicity /
teratogenicity", "Neurotoxicity", "Carcinogenicity",
        "Toxicity to reproduction", "Dermal
Absorption", "Specific investigations: other studies",
        "Immunotoxicity", "Exposure related
observations in humans: other data",
        "Toxicity to reproduction: other
studies", "Epidemiological data",
        "Sensitisation data (humans)",
        "Endocrine disrupter mammalian
screening - in vivo (level 3)"))
    if(length(subhead_rows)==0){head<-append(head[1],head)}
    else{if(subhead_rows[1] != 2){head<-append(head[1],head)}}
    subhead_rows<-which(head[2:length(head)] %in% c("Basic
toxicokinetics", "Acute toxicity: oral",
        "Acute toxicity: inhalation", "Acute
toxicity: dermal", "Acute toxicity: other routes",
        "Skin irritation / corrosion", "Eye
irritation",
        "Skin sensitisation", "Respiratory
sensitisation", "Repeated dose toxicity: oral",
        "Repeated dose toxicity:
inhalation", "Repeated dose toxicity: dermal",
        "Genetic toxicity in vitro", "Genetic
toxicity in vivo", "Dermal absorption",
        "Developmental toxicity /
teratogenicity", "Neurotoxicity", "Carcinogenicity",
        "Toxicity to reproduction", "Dermal
Absorption", "Specific investigations: other studies",
        "Immunotoxicity", "Exposure related
observations in humans: other data",
        "Toxicity to reproduction: other
studies", "Epidemiological data",
        "Sensitisation data (humans)",
        "Endocrine disrupter mammalian
screening - in vivo (level 3)"))

    subheadstop<-c()
    for(b in 1:(length(subhead_rows))) {
        if(b<length(subhead_rows)) {subheadstop[b] <- subhead_rows[b + 1] -
1}
        else{subheadstop[b]<-length(head)}

    ##loop 3 - subheaders
    if(length(subhead_rows) != 0){
    for(b in 1:length(subhead_rows)){
        subhead<-head[subhead_rows[b]:subheadstop[b]]
        neuro<-0

```

```

genovitro<-0
if(subhead[2]=="Neurotoxicity"){neuro<-1}
if(subhead[2]=="Genetic toxicity in vitro"){genovitro<-1}

#Identify record sections, start and stop
record_rows<-which(subhead=="ENDPOINT_STUDY_RECORD:")

recordstop<-c()
for(c in 1:(length(record_rows))) {
  if(c<length(record_rows)) {recordstop[c] <- record_rows[c + 1] -
1}
  else{recordstop[c]<-length(subhead)}}

##loop 4 - records
for(c in 1:length(record_rows)){
  record<-subhead[record_rows[c]:recordstop[c]]
  read<-0
  vitro<-0
  hum<-0
  anim<-0
  tox<-0
  half<-0
  max<-0
  met1<-0
  abs<-0
  met2<-0
  met3<-0
  iv<-0
  ip<-0
  dose<-0
  noael<-0
  end<-0
  mech<-0
  update<-matrix(0,1,ncol=4)
  update<-cbind(cas,subhead[2],update)

  #Identify sections and subsections, start and stop
  sect_rows<-which(record %in% c("Administrative data","Data
source","Materials and methods",
                                "Test material","In vivo test
system","Test animals","Administration / exposure",
                                "Method","Results and
discussion","Toxicokinetic / pharmacokinetic studies",
                                "Metabolite characterisation
studies","Endpoint","Type of information","Adequacy of study",
                                "Reliability","Reason / purpose for
cross-reference","Reference",
                                "Test guideline","Principles of
method if other than guideline",
                                "GLP compliance","Test type","Limit
test","Type of study","Type of assay",
                                "Test material
information","Radiolabelling","Species","Strain","Sex",

```

```

                                "Route of administration","Type of
inhalation exposure","Type of coverage",
                                "Duration of exposure","Duration of
treatment / exposure","Key result","Toxicokinetic parameters",
                                "Metabolites
identified","Objective","Objective of study","Main ADME results","Dose
descriptor",
                                "Effect level","Based
on","Generation"))

```

```

sectstop<-c()
for(d in 1:(length(sect_rows))) {
  if(d<length(sect_rows)) {sectstop[d] <- sect_rows[d + 1] - 1}
  else{sectstop[d]<-length(record)}}

##loop 5 - sections
for(d in 1:length(sect_rows)){
  #For each record, scan each section to identify key points
  sect<-record[sect_rows[d]:sectstop[d]]
  for(e in 2:length(sect)){
    if(sect[1] %in% c("Type of information","Reason / purpose for
cross-reference")){
      if(grepl("read-across",sect[e]) |
grepl("(Q) SAR",sect[e])){read<-1}}
    if(sect[1]=="Endpoint"){
      if(grepl("in vitro",sect[e])){vitro<-1}}
    if(sect[1]=="Species"){
      if(grepl("human",sect[e])){hum<-1}
      else{anim<-1}}
    if(sect[1] %in% c("Endpoint","Objective")){
      if(grepl("toxicokinetics",sect[e])){tox<-1}}
    if(sect[1]=="Toxicokinetic parameters"){
      if(grepl("half-life",sect[e])){half<-1}}
    if(sect[1]=="Toxicokinetic parameters"){
      if(grepl("Cmax",sect[e]) | grepl("Tmax",sect[e]) |
grepl("AUC",sect[e]) |
      grepl("transfer",sect[e]) | grepl("absorption",sect[e])
| grepl("distribution",sect[e]) |
      grepl("excretion",sect[e])){max<-1}}
    if(sect[1]=="Toxicokinetic parameters"){
      if(grepl("metabolism",sect[e]) |
grepl("metabolite",sect[e])){met1<-1}}
    if(sect[1] %in% c("Objective of study","Main ADME results")){
      if(grepl("absorption",sect[e]) |
grepl("distribution",sect[e]) | grepl("excretion",sect[e])){abs<-1}}
    if(sect[1] %in% c("Objective of study","Main ADME results")){
      if(grepl("metabolism",sect[e])){met2<-1}}
    if(sect[1]=="Metabolites identified"){
      if(grepl("yes",sect[e])){met3<-1}}
    if(sect[1]=="Route of administration"){
      if(grepl("intravenous",sect[e]) |
grepl("intramuscular",sect[e])){iv<-1}}
    if(sect[1]=="Route of administration"){
      if(grepl("intraperitoneal",sect[e])){ip<-1}}

```

```

        if(sect[1]=="Dose descriptor"){dose<-1}
        if(sect[1]=="Dose descriptor"){
            if(grepl("NOAEL",sect[2]) | grepl("LOAEL",sect[2]) |
grepl("NOAEC",sect[2]) |
            grepl("LOAEC",sect[2]) | grepl("NOEL",sect[2]) |
grepl("LOEL",sect[2])){noael<-1}}
        if(sect[1]=="Endpoint"){
            if(grepl("endocrine",sect[e])){end<-1}}
        if(sect[1]=="Endpoint"){
            if(grepl("mechanistic",sect[e])){mech<-1}}
    }
} #end loop 5

if(head[1]=="Toxicokinetics, metabolism and distribution"){
    if(vitro==1){update[,c(23,28)]<-1}
    else{
        if(hum==1 & tox==1 & half==1){update[,c(23,24,26)]<-1}
        if(hum==1 & tox==1 & max==1){update[,c(23,24)]<-1}
        if(hum==1 & tox==1 & met1==1){update[,c(23,26)]<-1}
        if(hum==1 & abs==1){update[,c(23,24)]<-1}
        if(hum==1 & met2==1){update[,c(23,26)]<-1}
        if(hum==1 & met3==1){update[,c(23,26)]<-1}
        if(anim==1 & tox==1 & half==1){update[,c(23,25,27)]<-1}
        if(anim==1 & tox==1 & max==1){update[,c(23,25)]<-1}
        if(anim==1 & tox==1 & met1==1){update[,c(23,27)]<-1}
        if(anim==1 & abs==1){update[,c(23,25)]<-1}
        if(anim==1 & met2==1){update[,c(23,27)]<-1}
        if(anim==1 & met3==1){update[,c(23,27)]<-1}}
    if(head[1]=="Acute Toxicity"){update[,c(3,4)]<-1}
    if(dose==1){update[,c(54,55)]<-1}}
    if(head[1]=="Irritation / corrosion"){update[,c(3,10)]<-1}
    if(head[1]=="Sensitisation"){
        if(hum==1){update[,c(13,21)]<-1}
        if(anim==1){update[,c(3,11)]<-1}}
    if(head[1]=="Genetic toxicity"){
        if(genovitro==1){update[,c(3,8)]<-1}
        else{if(hum==1){update[,c(13,18)]<-1}
            if(anim==1){update[,c(3,8)]<-1}}}
    if(head[1] %in% c("Toxicity to reproduction","Toxicity to
reproduction: other studies",
                    "Developmental toxicity / teratogenicity")){
        if(hum==1){update[,c(13,19)]<-1}
        if(anim==1){update[,c(3,9)]<-1}
        if(noael==1){update[,c(54,59)]<-1}}
    if(neuro==1){
        if(hum==1){update[,c(13,16)]<-1}
        if(anim==1){update[,c(3,6)]<-1}
        if(dose==1){update[,c(54,57)]<-1}}
    if(head[1]=="Carcinogenicity"){
        if(hum==1){update[,c(13,17)]<-1}
        if(anim==1){update[,c(3,7)]<-1}
        if(noael==1){update[,c(54,56)]<-1}}
    if(head[1]=="Repeated dose toxicity"){update[,c(3,5)]<-1}
    if(noael==1){update[,c(54,56)]<-1}}

```

```

        if(end==1){update[,c(3,12)]<-1
          if(noael==1){update[,c(54,61)]<-1}}
        if(mech==1){update[,31]<-1}
        if(subhead[2]=="Sensitisation data (humans)") {update[,c(13,21)]<-
1}

        if(read==1){
          if(head[1]=="Toxicokinetics, metabolism and
distribution") {update[,c(23,30)]<-1
            update[,c(3:22,24:28,30:(ncolt-2))]<-0}
            else{update[,34]<-1
              update[,35]<-max(update[,4],update[,14])
              update[,36]<-max(update[,5],update[,15])
              update[,37]<-max(update[,6],update[,16])
              update[,38]<-max(update[,7],update[,17])
              update[,39]<-max(update[,8],update[,18])
              update[,40]<-max(update[,9],update[,19])
              update[,41]<-max(update[,10],update[,20])
              update[,42]<-max(update[,11],update[,21])
              update[,43]<-max(update[,12],update[,21])
              update[,3:33]<-0
              update[,44:(ncolt-2)]<-0}}
          if((head[1]=="Genetic toxicity" & iv==1) | (head[1] != "Genetic
toxicity" & (iv==1 | ip==1))){
            update[,3:(ncolt-2)]<-0}

          #add to row of skeleton table
          all[row,3:(ncolt-2)]<-all[row,3:(ncolt-
2)]+as.numeric(update[1,3:(ncolt-2)])
        } #end loop 4

      } #end subheader loop
    }
  } #end if within loop 2
} #end header loop
}

}
}
} #end file loop

all<-cbind(all[,1:2],filter(tlist,!is.na(tlist$CAS))[,3:4],all[,3:(ncolt-
2)])
all<-bind_rows(filter(tlist,is.na(tlist$CAS)),all)
all<-all[order(all$CID,all$CAS,na.last=FALSE),]

wb <- loadWorkbook(outfile)
writeData(wb, sheet = "IUCLID", all[1:ncolt], colNames = F,startRow=3)
saveWorkbook(wb,outfile,overwrite = T)

```