Parsing Nist SDS-65.pdf

https://srdata.nist.gov/solubility/IUPAC/SDS-65/SDS-65.pdf

In [1]:
```
import pandas as pd
```

In [2]:
```
import fitz  # this is pymupdf
```

In [3]:
```
import re
```

In [4]:
```
import numpy as np
```

In [5]:
```
import os
```

In [6]:
```
import urllib
```

In [7]:
```
import io
```

In [8]:
```
from PIL import Image
```

Open book

In [9]:
```
bookfile = "SDS-65"
```

In [10]:
```
pages = fitz.open(f'Nist/{bookfile}.pdf')
```

Set parameters

In [11]:
```
charperline = 110
lineperpage = 200
```

Process page function

In [12]:
```
def processPage(page):
    words = page.getText("words")
    # words = sorted(words, key=lambda element: (int(element[3]), element[0]))
    # Calculate margins
    startwidth = 1000
    stopwidth = 0
    startheight = 1000
    stopheight = 0
    for word in words:
        if startwidth > word[0]:
            startwidth = word[0]
        if stopwidth < word[2]:
            stopwidth = word[2]
        if startheight > word[1]:
            startheight = word[1]
        if stopheight < word[3]:
            stopheight = word[3]
    cw = (stopwidth-startwidth)/charperline
    lh = (stopheight-startheight)/lineperpage
    lines = []
    for word in words:
        # y, h
        wyh = {word[3]:word[3]-word[1]}
```

```
        if wyh not in lines:
            lines.append(wyh)
    lines1 = []
    for line in lines:
        key = list(line.keys())[0]
        lines1.append((key,line[key]))
    lines1 = sorted(lines1, key=lambda elt: (elt[0], elt[1]))
    lines2 = []
    line = 0
    i = 0
    for line1 in lines1:
        y = line1[0]
        h = line1[1]
        if line == 0:
            line = 1
        else:
            if  abs(y-lasty) > (h+lasth)/8 and abs(y-h-lasty+lasth)  > (h+lasth)/8:
                line +=1
        elt = list(line1)
        elt.append(line)
        elt = tuple(elt)
        lines2.append(elt)
        lasty = y
        lasth = h
        i += 1
    lines = {}
    for line in lines2:
        lines.update({line[0]:line[2]})
    magwords = []
    for word in words:
        word1 = list(word)
        word1.append(lines[word[3]])
        magwords.append(word1)
    magwords = sorted(magwords, key=lambda element: (int(element[8]), element[0]))
    i = 0
    ptext = ""
    lasth = 1000
    for word in magwords:
        line = word[8]
        h = word[3]-word[1]
        y = word[3]
        w = word[2]-word[0]
        x = word[0]
        aword = word[4]
        if i == 0:
                nbchar = int((x-startwidth) / cw)
                ws = " " * nbchar
                buf = (ws + aword + " ")
        else:
            if line != lastline:
                r = 1
                if y - lasty > (h+lasth):
                    r = int(2*(y-lasty)/(h+lasth))
                ptext += buf + "\n" * r
                nbchar = int((x-startwidth) / cw)
                ws = " " * nbchar
                buf = (ws + aword + " ")
            else:
                nbchar = int((x-startwidth) / cw) - len(buf)
                ws = " " * (nbchar-1)
                buf += ( ws + aword +  " ")
        lasth = h
        lasty = y
        lastline = line
        i += 1
    ptext += buf
    ptext = ptext + '\n'
    # print(ptext)
    return ptext
```

Print a pages according to various method

In [13]:

```
startpage = 47
endpage = 47
```

Get the page image

In [14]:

```
zoom_x = 1.0 # horizontal zoom
zomm_y = 1.0  # vertical zoom
mat = fitz.Matrix(zoom_x, zomm_y)  # zoom factor 2 in each dimension
pix = pages[startpage].getPixmap(matrix = mat)  # use 'mat' instead of the identity matrix
```

In [15]:

```
Image.open(io.BytesIO(pix.getImageData()))
```

Out[15]:

COMPONENTS:	EVALUATOR:
(1) Copper(I) Chloride; CuCl; [7758-89-6] (2) Water; H_2O; [7732-18-5]	J. J. Fritz Department of Chemistry The Pennsylvania State University June, 1991

CRITICAL EVALUATION:

compounds formed were all relatively unstable and not water-soluble; however, water-soluble derivatives of any of the unsaturated compounds should promote solubility of CuCl, as observed with the unsaturated alcohols and organic acids discussed earlier.

PHASE DIAGRAMS OF SYSTEMS INVOLVING CuCl

A number of the investigations of the solubility of CuCl in aqueous chlorides contain either phase diagrams or the information needed to construct a phase diagram for the system studied. For some systems there has been only one such report. For those where phase diagram data are available in more than one publication, the various reports disagree. Because of this situation, no actual phase diagrams are included in this section. The available data (most of which is contained in the Compilations) will be discussed below.

$CuCl-HCl-H_2O$

The only phase diagram data for this system reported in the literature are those of Morosov and Ustanishkova[31], who measured the solubility of CuCl in aqueous HCl up to concentrations obtained when the solution was in equilibrium with gaseous HCl at atmospheric pressure (these are the last entries for each temperature in the Compilations of their data). (Note: Their data were presented in a different graphical form by Chaltykyan[15].) They gave the mass percentages of HCl and CuCl in these solutions; these are given in Table 2, along with the corresponding molalities calculated from them.

Table 2. Composition of Aqueous Solutions in Equilibrium with Solid CuCl and Gaseous HCl at Atmospheric Pressure

T/K	$100w_{CuCl}$	$100w_{HCl}$	$\frac{m_{CuCl}}{\text{mol kg}^{-1}}$	$\frac{m_{HCl}}{\text{mol kg}^{-1}}$
273	19.02	34.7	4.15	20.6
298	24.9	30.	5.58	18.2
323	26.0	27.5	5.65	16.2
353	29.13	23.25	6.18	13.4
373	30.0	19.27	5.97	10.4

Their graphical data also give the mass percentage of HCl for solutions under atmospheric pressure of HCl in the absence of dissolved CuCl. These data are given in Table 3. At 273 K they give a single point for a solution saturated with HCl under atmospheric pressure but unsaturated in CuCl. This occurs at 10 per cent CuCl, 39 per cent HCl (1.98 and 21.0 mol kg^{-1}, respectively).

Table 3. Composition of Solutions in Equilibrium with Gaseous HCl at Atmospheric Pressure in the Absence of Dissolved CuCl

T/K	$100w_{HCl}$	$\frac{m_{HCl}}{\text{mol kg}^{-1}}$
273	44.2	21.7
298	40.0	18.3
323	36.7	15.9
353	32.4	13.1
373	29.0	11.2

Comparison of the two tables indicates the small extent to which saturation with CuCl decreases the solubility of HCl in water at atmospheric pressure. Abundant data are available illustrating the same sort of effect in other systems involving CuCl with a soluble chloride.

Use getText

In [16]:

```
print(pages[startpage].getText())
```

```
COMPONENTS:
(1) Copper(I) Chloride; CuCI; [7758-89-6]
(2) Water; H20; [7732-18-5]
EVALUATOR:
J. J. FRITZ
Department of Chemistry
The Pennsylvania State University
June, 1991
27
CRITICAL EVALUATION:
compounds formed were all relatively unstable and not water-soluble; however, water-soluble
derivatives of any of the unsaturated compounds should promote solubility of CuCI, as
```

observed with the unsaturated alcohols and organic acids discussed earlier.

PHASE DIAGRAMS OF SYSTEMS INVOLVING CuCl

A number of the investigations of the solubility of CuCI in aqueous chlorides contain either phase diagrams or the information needed to construct a phase diagram for the system studied. For some systems there has been only one such report. For those where phase diagram data are available in more than one publication, the various reports disagree. Because of this situation, no actual phase diagrams are included in this section. The available data (most of which is contained in the Compilations) will be discussed below.

CuCl-HCl-H20

The only phase diafram data for this system reported in the literature are those of Morosov and Ustanishkova[3], who measured the solubility of CuCI in aqueous HCI up to concentrations obtained when the solution was in equilibrium with gaseous BCI at atmospheric pressure (these are the last entries for each temperature in the Compilations of their data). (Note: Their data were presented in a different graphical form by Chaltykyan[15].) They gave the mass percentages of HCI and CuGI in these solutions; these are given in Table 2, along with the corresponding molalities calculated from them.

Table 2. Composition of Aqueous Solutions in Equilibrium with Solid CuCI and Gaseous HCI at Atmospheric Pressure

T/K	100WCuCI	100WHCl	mCuCI mol kg- 1	mHCI mol kg- 1
273	19.02	34.7	4.15	20.6
298	24.9	30.	5.58	18.2
323	26.0	27.5	5.65	16.2
353	29.13	23.25	6.18	13.4
373	30.0	19.27	5.97	10.4

Their graphical data also give the mass percentage of HCI for solutions under atmospheric pressure of HCI in the absence of dissolved CuCI. These data are given in Table 3. At 273 K they give a single point for a solution saturated with HCI under atmospheric pressure but unsaturated in CuCI. This occurs at 10 per cent CuCI, 39 per cent HCI (1.98 and 21.0 mol kg-I, respectively).

Table 3. Composition of Solutions in Equilibrium with Gaseous HCI at Atmospheric Pressure in the Absence of Dissolved CuCl

T/K	100WHCI	mHCI mol kg- 1
273	44.2	21.7
298	40.0	18.3
323	36.7	15.9
353	32.4	13.1
373	29.0	11.2

Comparison of the two tables indicates the small extent to which saturation with CuCI decreases the solubility of HCI in water at atmospheric pressure. Abundant data are available illustrating the same sort of effect in other systems involving CuCI with a soluble chloride.

Use processPage

In [17]:
```
print(processPage(pages[startpage]))
```

```
                                                                                  27
COMPONENTS:                                            EVALUATOR:
 (1) Copper(I)   Chloride;  CuCI;   [7758-89-6]         J. J. FRITZ
 (2) Water;  H20;   [7732-18-5]                         Department     of Chemistry
                                                         The  Pennsylvania    State  University
                                                         June, 1991
CRITICAL     EVALUATION:
 compounds   formed   were  all relatively unstable   and  not water-soluble;   however,   water-soluble
 derivatives of  any  of  the  unsaturated    compounds    should   promote    solubility  of  CuCI,    as
 observed  with  the  unsaturated   alcohols  and  organic   acids  discussed  earlier.
               PHASE     DIAGRAMS          OF   SYSTEMS        INVOLVING         CuCl
 A number   of the  investigations   of the solubility  of CuCI   in aqueous   chlorides  contain  either
 phase diagrams   or  the information   needed   to construct   a phase  diagram   for the system   stud-
 ied. For some  systems   there  has  been  only  one  such  report.  For  those  where  phase   diagram
 data are  available  in more   than  one  publication,   the  various  reports   disagree.   Because   of
 this situation, no  actual   phase  diagrams    are included   in  this section.   The   available  data
 (most of which   is contained   in the  Compilations)    will be discussed    below.
 CuCl-HCl-H20
 The only  phase  diafram    data for this system   reported   in the literature  are those  of Morosov
 and Ustanishkova3     , who   measured    the  solubility of  CuCI   in aqueous    HCI   up  to concen-
 trations obtained   when   the  solution  was   in equilibrium    with  gaseous   BCI   at atmospheric
 pressure (these  are  the last entries for  each  temperature    in the Compilations    of their data).
 (Note: Their  data  were  presented   in a different graphical   form  by  Chaltykyan15.)    They   gave
 the mass  percentages   of  HCI  and  CuGI   in these  solutions;  these  are given  in  Table  2, along
 with the  corresponding    molalities  calculated  from   them.
      Table  2. Composition        of  Aqueous      Solutions     in  Equilibrium      with    Solid
                CuCI   and   Gaseous     HCI    at Atmospheric       Pressure
                                                                mCuCI         mHCI
                         T/K     100WCuCI     100WHCl      mol  kg-   1    mol  kg-  1
                          273      19.02        34.7          4.15          20.6
                          298      24.9         30.          5.58          18.2
                          323      26.0         27.5          5.65          16.2
                          353      29.13        23.25         6.18          13.4
                          373      30.0         19.27         5.97          10.4
 Their graphical   data  also give  the mass   percentage   of  HCI  for solutions  under   atmospheric
 pressure of  HCI   in the  absence   of dissolved   CuCI.    These   data  are  given  in Table   3.  At
 273 K they  give  a single point  for a solution  saturated   with  HCI  under   atmospheric    pressure
 but unsaturated    in  CuCI.    This   occurs  at  10  per  cent  CuCI,    39  per  cent  HCI   (1.98  and
 21.0 mol  kg-I,  respectively).
      Table  3. Composition       of  Solutions     in Equilibrium       with   Gaseous     HCI    at
                Atmospheric      Pressure     in  the  Absence     of  Dissolved     CuCl
                                                                 mHCI
                                       T/K    100WHCI      mol  kg-  1
                                        273      44.2         21.7
                                       298      40.0         18.3
                                       323      36.7         15.9
                                       353      32.4         13.1
                                        373      29.0         11.2
 Comparison    of the  two   tables  indicates  the  small  extent   to  which  saturation    with  CuCI
 decreases the solubility  of HCI  in water  at atmospheric   pressure.   Abundant    data  are available
 illustrating the same   sort of effect in other  systems   involving   CuCI   with  a soluble  chloride.
```

Do the Book now

In [18]:
```
startpage = 0
endpage = 311
```

In [19]:
```
%%time
text = ""
for i in range(startpage,endpage+1):
    text += processPage(pages[i])
```

```
Wall time: 2.5 s
```

Save as text

In [20]:
```
file=f'Nist/{bookfile}.txt'
with open(file, 'w', encoding='utf8') as filetowrite:
    filetowrite.write(text)
```

```
In [ ]:
```