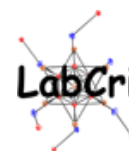


Laboratório de Cristalografia

Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - Pampulha
Belo Horizonte (MG) Brasil CEP: 31270-901 Cx: 702
TEL: (+55 31) - 3409 6600 / 3409 5632 FAX: (+55 31) 3409 5600



Intensive Crystallography Course 2011



Data Mining

Alexandre Melo, Vítor Macaroun, João O. S. Mendes &
Carlos B. Pinheiro

Practical Exercise



1. General Objectives

On the next pages you will find a short description of the major steps leading to data mining regarding to retrieve structural information on chemical compounds and proteins. The idea is to present you some of the databases in which structural information can be retrieved.

2. Databases and information research

The first step for solving a structure (any structure) is to collect all potentially available information that can help in this task. Either information about the same compound or other chemically related compound could be useful. Hopefully there already exist many different databases that can be exploited depending on the kind of compound under investigation. For example:

- **Web of Science** - From the university computers use *www.capes.gov.br* → *Web of Science*. It is an important tool to search for scientific articles, books and citation index. In particular it allows using citations for query.
- **Science Direct** - *http://www.sciencedirect.com*. Like Web of Science, it is an important tool to search for scientific ar(any structure) is to collect all potentially available information that can help in this task. Either information about the same compound or other chemically related compound could be useful. Hopefully there already exist many different databases that can be exploited depending on the
- **Crystal Structure Databases** - Cambridge Crystallographic Data Center (CCDC), Inorganic Crystal Structure Database (ICSD), Protein Data Base (PDB), American Mineralogist Crystal Structure Database, IUCr-Crystallography Journals Online, etc.

Crystal Structure Databases

These kinds of databases are of special interest for crystallography. A list of them can be found in the server of the IUCr (International Union of Crystallography). The address is: <http://www.iucr.org/data/index.html>. Among the important Crystallographic Databases some have to be highlighted.

Cambridge Structural Database: from the Cambridge Crystallographic Data Center (CCDC) specialized only in organic and metal-organic compounds. It contains more than 500.000 entries, and it is nowadays the biggest data bank for organic and metal-organic compounds. The structures describe there were determined by neutron or X-ray diffraction techniques. The access is limited and a password is needed. The CCDC gives also the possibility of downloading *Crystallographic Information File* - CIF - (<http://www.iucr.org/resources/cif>) and structural analysis files. More information at <http://www.ccdc.cam.ac.uk/>.

ICSD-Inorganic Crystal Structure Database: FIZ Karlsruhe provides the scientific and the industrial community with the world's largest database for completely identified inorganic crystal structures, ICSD, containing about 135,500 peer-reviewed data entries including their atomic coordinates dating back to 1913. It contains structures determined by neutron or X-ray diffraction techniques. More information at http://www.fiz-karlsruhe.de/icsd_home.html. The public access is limited (<http://icsd.fiz-karlsruhe.de/> DEMO ACCOUNT), however the PORTAL CAPES gives you full-unlimited access, see below.

PDB-Protein Data Bank: The PDB database contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. More than 64000 macromolecular structures determined by X-ray diffraction or RMN spectroscopy are registered. The access is free from <http://www.rcsb.org/pdb/home/home.do>.

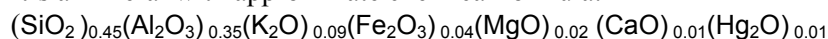
IUCr-Crystallography Journals Online: The IUCr server gives access to all articles published in its journals complemented with more information like CIF and HKL (experimental data) files. *ACTA A, B* and *C* are freely available from the PORTAL CAPES at <http://journals.iucr.org>.

American Mineralogist Crystal Structure Database: from the Mineralogical Society of America only for minerals. It is an example of small useful data bank. It is possible to look for mineral structures by mineral name or by author's name from the papers published at the "American Mineralogist". Free access via Internet at: www.geo.arizona.edu/AMS/

3. Examples

Mica (AMCSD)

It's a mineral with approximate chemical formula:



We will try at the "American Mineralogist Crystal Structure Database".

- Type "mica" in the "mineral" field.
- Click on "search"
- Discuss and analyze the results. Get familiar with the different options of this database.

TCoenCl (trans-dichloro-bis(ethylenediamine)-cobalt(iii) chloride (CCDC)

Important: write down and keep all data that you will find for the next compound! Later you will use this information to solve your structure. A draw of the molecular shape can be very useful.

We will look for the structure of TCoenCl at the "CCDC" database.

- Connect to CCDC.
 - In order to do that connect your computer in the *crystal6.fisica.ufmg.br* (150.164.15.150) server and use the account **XXXX**.
- Open a *Xterm* window and type "*cq &*" in the command line.
- In order to build a query in the Conquest click on the "*Build Queries*" tab.
- First, click on "Name/Class" and build a query with "*ethylenediamine*", as indicated in the figure below.

Name/Class (2) - New

Compound Name
ethylenediamine

Ignore non-alphabetic characters,
e.g. "butadiene" will match "buta-1,3-diene"

Find exact word,
e.g. "hydrazine" will not match "acetylhydrazine"

Add Replace Delete

Contains:

Chemical Class

CCDC Chemical Class is assigned to some categories of entries in the CSD, particularly when it would be difficult to locate these categories using a substructure or compound name search.

Note that the results of a search on Class may not be comprehensive since it is not always evident from a publication that an entry belongs to one of the specified categories.

Find entries classified as: ----not defined----

and: ----not defined----

Search Store Cancel Reset

- How many compounds do you find? Inspect some of them.
- If you obtain too many results try to reduce your research including more information. To do this, click on “*Build Queries*” tab and choose “*Elements*” to make a query with cobalt and chloride. Then, click “*Store*”. See the example in the figure below.

Elements (2) - New

Elements Required to be Present
Co Cl

Type in elements, e.g. C H Se
or Select from Table

Elements must be in

same molecule

same crystal structure

Other elements allowed in molecule/structure

Heaviest Permitted Element in Formula Unit
-- Not Set -- Select from Table

Search Store Cancel Reset

- Now click on “*Combine Queries*” tab to combine the information and to proceed with a search. How many compounds do you find? ? Inspect some of them

- Try to use unit cell information to find something. Use the data presented in the image below.

Unit Cell (2) - New

Do you want to search on the reduced cell?
You should search on reduced cell if you want to find structures which match a particular set of cell dimensions (a,b,c,alpha,beta,gamma)

Yes, do a reduced cell search No, do not do a reduced cell search

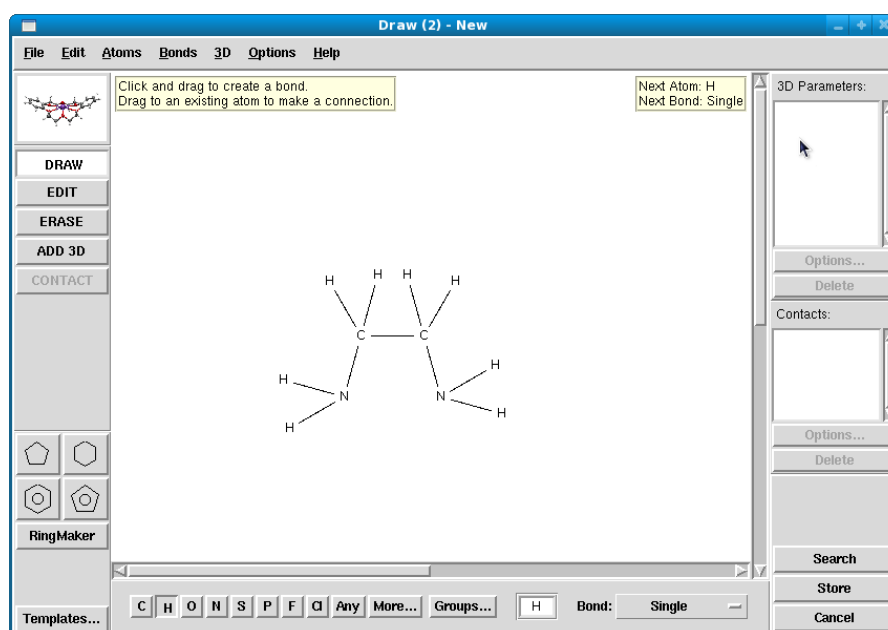
Tolerance % of longest cell dimension

Lattice Type

Cell Parameters

a (A)	=	<input type="text" value="6.3"/>	alpha (°)	=	<input type="text" value="90"/>
b (A)	=	<input type="text" value="9"/>	beta (°)	=	<input type="text" value="108"/>
c (A)	=	<input type="text" value="9.4"/>	gamma (°)	=	<input type="text" value="90"/>

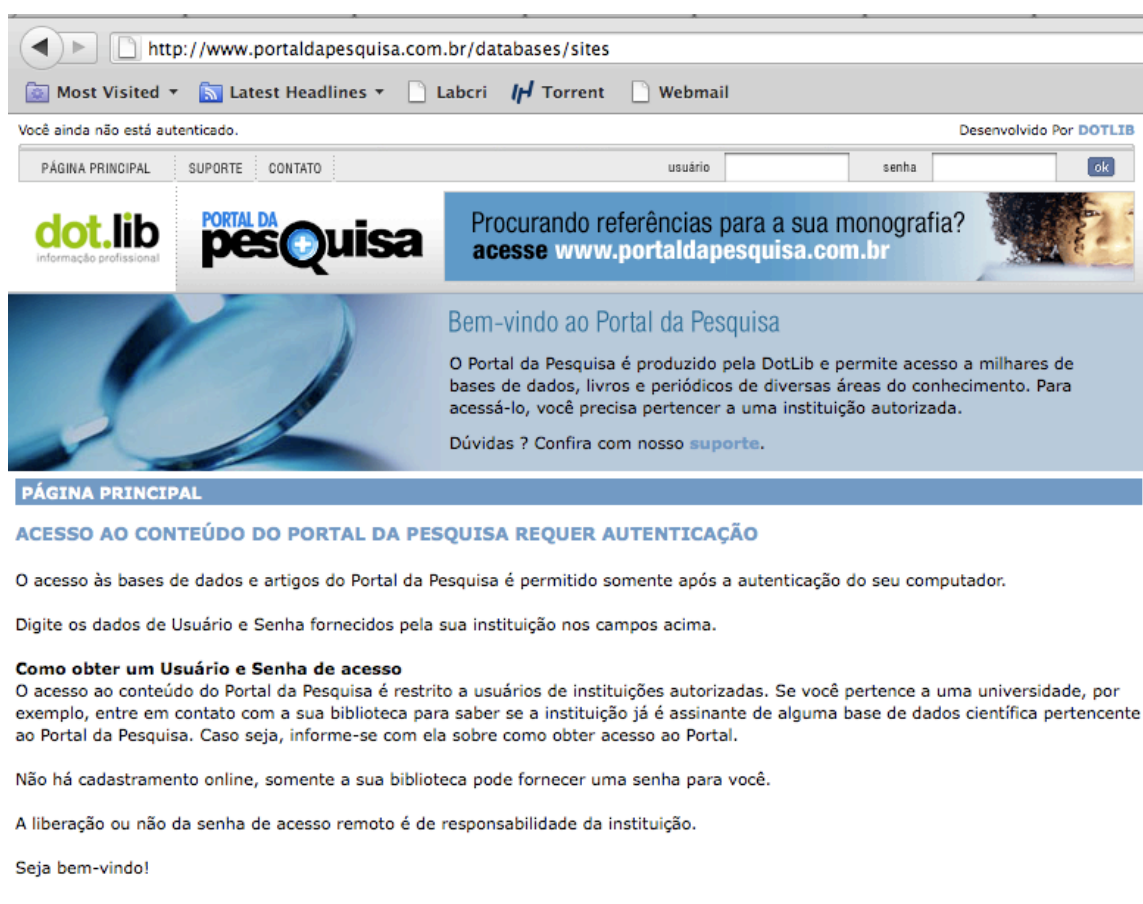
- Now click on “*Combine Queries*” tab to combine all information you have and to proceed with a search. How many compounds do you find?
- Finally, if you know something about the structure of compound, try to draw its diagram or part of it. Use the “*draw*” tab and play around. Try to draw the following diagram



- Now click on “*Combine Queries*” tab to combine all information you have and to proceed with a search. How many compounds do you find? Can you say for sure anything about the structure of the TCoenCl compound?

SiO₂ (ICSD PORTAL CAPES)

- Access the web site: <http://www.portaldapesquisa.com.br/databases/sites>.



The screenshot shows a web browser window with the URL <http://www.portaldapesquisa.com.br/databases/sites>. The page header includes navigation links for 'PÁGINA PRINCIPAL', 'SUPORTE', and 'CONTATO', along with a login form for 'usuário' and 'senha'. The main content area features a blue banner with the text 'Bem-vindo ao Portal da Pesquisa' and a message stating that access to the database requires authentication. Below this, there is a section titled 'ACESSO AO CONTEÚDO DO PORTAL DA PESQUISA REQUER AUTENTICAÇÃO' with instructions on how to obtain access, including contact with the library and the need for a password provided by the institution.

- Follow the instruction therein for obtaining the access.
- After log in, chose among the available datasets the
- [ICSD - INORGANIC CRYSTAL STRUCTURE DATABASE](#).

Portal da Pesquisa - Mozilla Firefox

http://www.portaldapesquisa.com.br/databases/sites

Portal da Pesquisa

Desenvolvido Por DOTLIB

ÁREA: Ciências Exatas e da Terra EDITORA: Todas Editorias

TÍTULO: pesq:##

BASES DE DADOS > CIÊNCIAS EXATAS E DA TERRA

- AMERICAN MINERALOGIST CRYSTAL STRUCTURE DATABASE
- CRYSTALLOGRAPHY OPEN DATABASE
- CRYSTMET
- ICSD - INORGANIC CRYSTAL STRUCTURE DATABASE**
- MINERALOGY DATABASE
- NUCLEIC ACID DATABASE
- PROTEIN DATA BANK

© 2003-2010 - DOTLIB - Todos os direitos reservados.

In the example below, all information concerning the SiO_2 compound will be retrieved from ICSD.

- Inside ICSD chose *Navigation* menu followed by *Chemistry* option.

ICSD - Mozilla Firefox

http://icsd.fz-karlsruhe.de/w10001.dotlib.com.br/

Welcome to ICSDWeb, IP authenticated (64.135.27.1), Dot Lib

Navigation

- Basic search & retrieve
- Advanced search & retrieve
- Bibliography
- Cell
- Chemistry**
- Symmetry
- Crystal Chemistry
- Structure Type
- Experimental Information
- DB Info
- Query Management
 - Load/Modify Queries
 - Save Queries
 - Delete Queries

Online Survey 2010

Please take some time and participate in our short online survey. Your feedback is very important and highly appreciated. Your responses will help us improve the ICSD according to your needs and to address any issues that you may have.

This survey should take less than 5 minutes to complete. Your responses will be kept confidential and will be used for internal evaluation only.

If you have any further comments or feedback, you are welcome to contact Stephan Rühl.

You may find the online survey [here](#).

Introduction to ICSD Web

Welcome to the ICSD Web version
You are now logged in to ICSD

Please click „Basic search & retrieve“ or „Advanced search & retrieve“ to enter your retrieval dialog.

On the left menu there are two different buttons. With "Personalize account" You can additionally registrate yourself with an extra Login ID and Password. This feature enables you to store personal queries etc.

After you personalized your account, please click the "Login" button to use your customized settings.

Current Content

The database now contains 135.500 entries.
At present, the ICSD contains

- 1.514 crystal structures of the elements
- 25.923 records for binary compounds
- 48.920 records for ternary compounds
- 49.525 records for quaternary and quinary compounds

Detailed information on the ICSD may be found in the [scientific manual](#) and on our [information flyer](#).

Revised data

Search Action

Run Query Save Query Clear Query

Search Summary

- Bibliography: -
- Cell: -
- Chemistry: -
- Symmetry: -
- Crystal Chemistry: -
- Structure Types: -
- Experimental Info: -
- DB Info: -

Query History

Number of queries: 0

javascript:submitData("", "de.fz.icsd.tiles.search.ChemistryQuickLayout", "Chemistry")

- In order to limit the number of answers of a specific query it is recommended to perform the search in the sub-menu *ADVANCED SEARCH & RETRIEVE* → *CHEMISTRY*.

The screenshot shows the ICSD website interface in Mozilla Firefox. The browser address bar displays the URL: <http://icsd.fiz-karlsruhe.de.w10001.dotlib.com.br/icsd/StartActionPath.do?sessionId=EDF7056CB24F63F958AD8844238ACD5>. The page title is "ICSD - Mozilla Firefox". The main content area is titled "Search Chemistry Quick Search mode" and includes a navigation menu on the left with options like "Basic search & retrieve", "Advanced search & retrieve", and "Query Management". The search form contains fields for Composition (e.g., Na Cl), Structural Formula (e.g., Pb (W O4)), Chemical Name, Mineral Name (e.g., Adamite), Mineral Group (e.g., Pyroxene), ANX Formula, Cryst. Comp., AB Formula, and Chem. Comp. There are also input fields for "Number of Elements" and "Number of Formula Units". Buttons for "Clear Chemistry Search" and "Count Chemistry Search" are visible at the bottom of the form. The right sidebar shows "Search Action" (Run Query, Save Query, Clear Query) and "Search Summary" (Bibliography, Cell, Chemistry, Symmetry, Crystal Chemistry, Structure Types, Experimental Info, DB Info). The footer includes "Legal Notices | Privacy Policy | Disclaimer | Copyright © FIZ Karlsruhe 2010" and "Version 1.3.0 (build 20101021-1203)".

The screenshot shows the Windows taskbar. The taskbar includes the Start button (Iniciar), several open application windows (ICSD - Mozilla..., ICSD_instrução...), and the system tray showing the time as 15:19.

- Inside the sub-menu *CHEMISTRY* select the option *VISUAL SEARCH MODE*.

The screenshot shows the ICSD website interface in Mozilla Firefox, now in "Search Chemistry Visual Search mode". The browser address bar displays the URL: <http://icsd.fiz-karlsruhe.de.w10001.dotlib.com.br/icsd/StartActionPath.do>. The main content area features a periodic table of elements. The search form is similar to the previous screenshot but includes a "Restrict total number of elements to selected number of elements" checkbox. The right sidebar shows "Search Action" (Run Query, Save Query, Clear Query) and "Search Summary" (Bibliography, Cell, Chemistry, Symmetry, Crystal Chemistry, Structure Types, Experimental Info, DB Info). The footer includes "Legal Notices | Privacy Policy | Disclaimer | Copyright © FIZ Karlsruhe 2010" and "Version 1.3.0 (build 20101021-1203)".

The screenshot shows the Windows taskbar. The taskbar includes the Start button (Iniciar), several open application windows (ICSD - Mozilla..., ICSD_instrução...), and the system tray showing the time as 15:19.

- In the periodic table select the elements of the compound SiO_2 you are looking for. They will be listed in the bottom of the screen. Fill options “*Co.(min)*” and “*Co.(max)*” with the number of atoms of the compound you are looking for. “*Co.(min)*” and “*Co.(max)*” should be different only if you doubt the compound composition and need to perform a more broad research.
- If you are sure about the composition of the material you are looking for, select “*Restrict total number of elements to selected number of elements*”, just below the list of elements.
- In the SEARCH ACTION menu (to-left) select RUN QUERY.
- After some seconds a list of compounds matching your options organized in chronologic order (smaller to biggest “*Coll. Code*”) will be displayed on the screen

The screenshot displays the ICSD List View Page in Mozilla Firefox. The browser address bar shows the URL: <http://icsd.fiz-karlsruhe.de/w10001.dotlib.com.br/viscalc/jsp/listView.action?SESSIONID=EDF7056CB24F63F95F8AD884423BACD5>. The page title is "FIZ Karlsruhe ICSD: List View Page - Mozilla Firefox".

The main content area shows a search results table with the following columns: Coll. Code, HMS, Struct. Form., Struct. Type, Title, Authors, and Reference. The results are sorted by Coll. Code in ascending order. The first few entries are:

Coll. Code	HMS	Struct. Form.	Struct. Type	Title	Authors	Reference
176	C 1 c 1	Si O2	SiO2(mS144)	Kristallstruktur des monoklinen Trif-Tridymite	Kato, K.; Nukui, A.	Acta Crystallographica B (24, 1968-38, 1982) (1976) 32, p2496-p2491
1109	C 1 c 1	Si O2	SiO2(mS144)	Silicon-oxygen bond lengths, bridging angles Si-O-Si and synthetic low tridymite	Baur, W.H.	Acta Crystallographica B (24, 1968-38, 1982) (1977) 33, p2615-p2619
9160	P 42/m n m	Si O2	TiO2(P6)	Rutile-type compounds. VI. Si O2, Ge O2 and a comparison with other rutile-type structures	Baur, W.H.; Khan, A.A.	Acta Crystallographica B (24, 1968-38, 1982) (1971) 27, p2133-p2139
10078	P 42/m n m	Si O2	TiO2(P6)	Single crystal analysis of the structure of stishovite	Sinclair, W.; Ringwood, A.E.	Nature (London) (1978) 272, p714-p715
16331	P 32.2 1	Si O2	Quartz,low	Structure determination of alpha-quartz up to 60°(hexagonal Pe	d'Amour, H.; Denner, W.; Schulz, H.	Acta Crystallographica B (24, 1968-38, 1982) (1979) 35, p550-p555
18112	C 1 2/c 1	Si O2	Coesite	Refinement of a coesite structure	Airaki, T.; Zoltai, T.	Zeitschrift fuer Kristallographie, Kristallgeometrie, Kristalphysik, Kristalchemie (-144, 1977) (1969) 129, p381-p387
27745	P 3 1 2 1	Si O2	Quartz,low	Crystal structure of neutron-irradiated alpha-quartz	Korneev, A.E.; Betokoneva, E.L.; Kolontsova, E.V.; Simonov, M.A.	Kristallografiya (1978) 23, p412-p413
30869	C 1 2/c 1	Si O2	Coesite	Ending the "P21/a coesite" discussion	Kirfel, A.; Will, G.	Zeitschrift fuer Kristallographie (149, 1979-) (1984) 167, p287-p291
31312	C 1 2/c 1	Si O2	Si O2	Re-examination of "P21/a coesite"	Sasaki, S.; Chen, H.K.; Prewitt, T.; Nakajima, Y.	Zeitschrift fuer Kristallographie (149, 1979-) (1983) 164, p67-p77
34867	C 1 c 1	Si O2	SiO2(mS144)	The superstructure of meteoritic low tridymite solved by computer simulation	Dolase, W.A.; Baur, W.H.	American Mineralogist (1976) 61, p971-p978

The page also includes navigation and quality filter options on the left side. The quality filter is set to "All Data".

- The output list can also be sorted as function of many other parameters according to the selection of the items in the top of the list.
 - Space group (“*HMS*”);
 - Structural formulae (“*Struct. Form.*”);
 - Structure type (“*Struct. Type.*”);
 - Publication title and/or authors;
 - Data quality (★ indicates high quality data).

TIP: sort the data according to quality. Quality in the ICSD is defined by the following parameters:

- Structure determination including refinement (in case of powder data including Rietveld refinement)
- Temperature factors given
- Pressure in the range 0.09 MPa to 0.11 MPa

- Temperature in the range 285 K to 300K
 - Standard deviation given for cell parameters
 - Any R-values.
- If you retrieve a high number of options that includes data of different quality, you can select in the option “*High Quality Data*” in the menu QUALITY FILTER.
 - Select the data that better fit to your research (box selection).
 - Select the option SHOW DETAILED VIEW and inspect the information. It will look like the figure below.

The screenshot shows the ICSD Detailed View page for Collection Code 176. The page is displayed in a Mozilla Firefox browser window. The browser's address bar shows the URL: <http://icsd.fiz-karlsruhe.de/w10001.dotlib.com.br/viscalc/jsp/sliderDetailed.action>. The page header includes the ICSD logo and navigation links. The main content area is titled "Detailed View" and shows "Entry 1 of 1". Below this, there is a "Summary" section for "Collection Code 176". The summary table includes the following data:

Summary		Collection Code 176	
Struct. formula	Si O2	Author	Kato, K.; Nului, A.
Space Group	C 1 c 1(9)	Title of Article	Kristallstruktur des monoklinen Tief-Tridymits
Unit Cell	18.4940(80) 4.991(2) 25.8320(80) 90. 117.75(2) 90.	Reference	Acta Crystallographica B (24.1968-38, 1982) (1976) 32, p2488-p2491
Cell Volume	2110.15 Å ³	Formula Units per Cell	48
Temperature	293.00 K (default)	Pressure	0.101325 MPa (default)
PDF-numbers	01-071-0032 18-1170	R-value	0.085
Remark	High Quality Data		

Below the summary, there are buttons for "Export CIF File" and "MyBaseFileName". The "Details" section is expanded, showing a list of options: Visualization, Chemistry, Published Crystal Structure Data, Standardized Crystal Structure Data, Distances & Angles, Bibliography, Experimental Information, Warnings & Comments, and Compare Published & Standardized Structure. At the bottom of the page, there are links for "Legal Notices", "Privacy Policy", "Disclaimer", and "Copyright © FIZ Karlsruhe 2010".

The screenshot shows a Windows taskbar. The taskbar includes the 'Concluído' status, the 'Iniciar' button, and two open windows: 'FIZ Karlsruhe ...' and 'ICSD_instrução...'. The system clock in the bottom right corner shows 15:25.

All the information about the specific compound within the database is displayed in summary page. Further information can be found in the DETAILS tab. You can also export the CIF ([Crystallographic Information File](#)) and/or visualize the structure and its diffraction pattern by selecting the VISUALIZATION option in the DETAILS tab.

FIZ Karlsruhe ICSD: Detailed View Collection Code 176 - Mozilla Firefox

http://icsd.fiz-karlsruhe.de/w10001.dotlib.com.br/viscalc/jsp/sliderDetailed.action

Portal da Pesquisa

Summary

Collection Code 176	
Struct. formula	Si O ₂
Space Group	C 1 c 1 (9)
Unit Cell	18.494(8) 4.991(2) 25.8320(8) 90 117.75(2) 90
Cell Volume	2110.15 Å ³
Temperature	293.00 K (default)
Pressure	0.101325 MPa (default)
PDF-numbers	01-071-0032 18-1170
R-value	0.085
Author	Kato, K.; Nukui, A.
Title of Article	Kristallstruktur des monoklinen Tief-Tridymits
Reference	Acta Crystallographica B (24, 1968-38, 1962) (1976) 32, p2486-p2491
Warnings & Comments	0 Warnings / 2 Comments

Remark: High Quality Data

Export CIF File MyBaseFileName Feedback to the ICSD Editor

Details

Expand All Collapse All

Visualization

Published Crystal Structure

Si O₂ - 1976 Kato, K., Nukui

C 1 c 1

a=18,494 Å

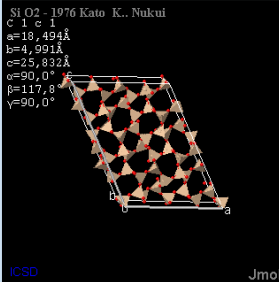
b=4,991 Å

c=25,832 Å

α=90,0°

β=117,8°

γ=90,0°



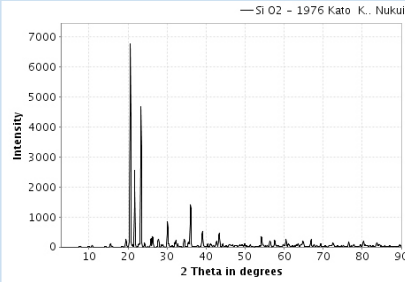
ICSD Jmol

Display in Window

Configure Structure Display

Powder Pattern of ICSD Coll.Code: 176

— Si O₂ - 1976 Kato, K., Nukui



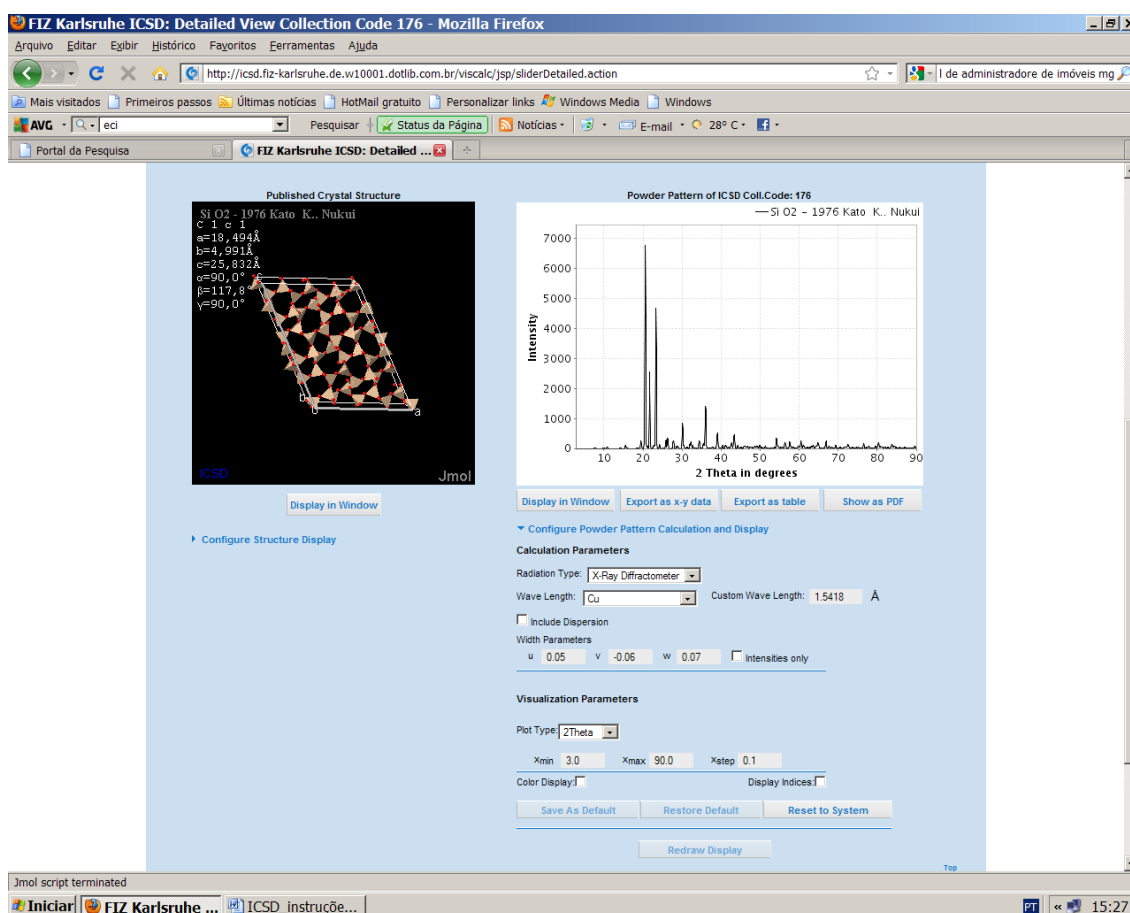
Intensity

2 Theta in degrees

Display in Window Export as x-y data Export as table Show as PDF

Configure Powder Pattern Calculation and Display

When visualizing the diffraction pattern within this option the diffraction pattern can be modified (wavelength, angular step, etc.) using the option CONFIGURE POWDER PATTERN CALCULATION AND DISPLAY.



The diffraction data can be exported in a file containing angular position \times intensity (NAME.DAT). This diffraction pattern might be useful for comparing experimental and database data. Try to play around to get the file in a format suitable for working with the graphical program of your preference.

4. Data mining: practical exercise

Important: write down and keep all data that you will find for the next compounds! Later you will use this information to solve your structure. A draw of the molecular shape can be very useful.

- A.** In CCDC look for and retrieve all information you can find about sucrose/saccharose.
- B.** In ICSD look for and retrieve all information you can find about Magnetite Fe₃O₄.
- C.** In ICSD look for and retrieve all information you can find about the salt NaCl.