

Automation of Institute Code



VIDYA DEVI

IAEA VIENNA



JAGJIT SINGH

UIET, PANJAB UNIVERSITY

CHANDIGARH, INDIA

The purpose of the presentation here is just to give a basic idea to extract some fields from the PDF file of the research paper for EXFOR compilation.

Objectives

Create a base prototype using Python programming language

- To read the data from the PDF of research paper
- To scan the data for the following keywords
 - DOI No., Authors Name, Institutes name
- To compare and search the institute names in the TRANS Dictionary
- To write the Institute Codes in the EXFOR format

Methodology

Institute name could be extracted in two ways:

- **OPTION 1:** Scanning DOI number from the PDF or as input from the user and parsing the research paper website
- **OPTION 2:** Scanning the whole PDF and search the required data from the text

Methodology

The **OPTION 2** of scanning the whole PDF looks more promising to even scan the FACILITY within the research paper in future.

We adopted **OPTION 2** to perform the following tasks:

- Scanned the pdf using python inbuilt function **PyPdf2**.
- Scanned for the DOI number to obtain the Bibliography.
- Scanned the text for the institute names.

Methodology for scanning Institute

TEXT EXTRACTION from PDF: The code truncate the text up to Abstract to limit the search.

- First it search country name and corresponding city within the text, say,

Country_Name = CHINA, City_Name= LINFEN

- The keywords like Department, INST, UNIV, NAT, **City_Name**, **Country_Name** etc. will be used to search the lines containing institute name within the text.

Methodology for scanning Institute

- After extracting institute name it create keywords from it
- For example, if institute name is
“Institute of Modern Physics, Shanxi Normal University, Linfen, China”
Then the keywords, say **KEY1**, will be
KEY1 :INST, MODERN, PHY, SHANXI, NORMAL, UNIV, LINFEN

Methodology for scanning Institute

Text extraction from TRANS DICTIONARY: In Trans Dictionary it extract lines using Country_Name.

For example, if Country Name is CHINA,

It search for the Country code, i.e., Country_Code = 3CPR

- It extract all the text lines from TRANS Dictionary beginning with 3CPR, such as,

3CPRSHN	(Shaanxi Normal Univ., Xian)	3000000301095
3CPRSIU	(Sichuan Univ., Chengdu)	3000000301097
3CPRSNU	(Shanxi Normal University, Linfen)	3000000301098

Methodology for scanning Institute

The code will search for keywords **KEY1** within the extracted lines of TRANS Dictionary, i.e.,


3CPRSNU (Shanxi Normal University, Linfen)

If it could not find the code within TRANS Dictionary, it will return

Country_Code = 3CPRCPR

#: We could also extract keywords from TRANS Dictionary to compare the institute text

Code has been run on four research papers from three different journals

- Chinese Journal of Physics
 - Physical Review C
 - Nuclear Physics A
- 

Example 1

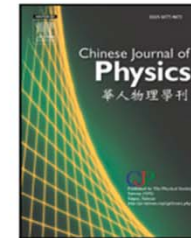
Chinese Journal of Physics 77 (2022) 1145–1155



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Chinese Journal of Physics

journal homepage: www.elsevier.com/locate/cjph



Measurement of cross sections for charge pickup by ^{40}Ar on elemental targets at 500 MeV/n

Su-Hua Zheng^{a,b}, Hai-Rui Duan^a, Jing-Ya Wu^a, Jun-Sheng Li^a, Satoshi Kodaira^c, Dong-Hai Zhang^{a,*}

^a Institute of Modern Physics, Shanxi Normal University, Linfen 041004, China

^b Department of Science, Taiyuan Institute of Technology, Taiyuan 03008, China

^c National Institute of Radiological Sciences, National Institutes for Quantum and Radiological Science and Technology, 4-9-1 Anagawa, Inage-ku, Chiba, 263-8555, Japan

ARTICLE INFO

Keywords:

Charge pickup reaction

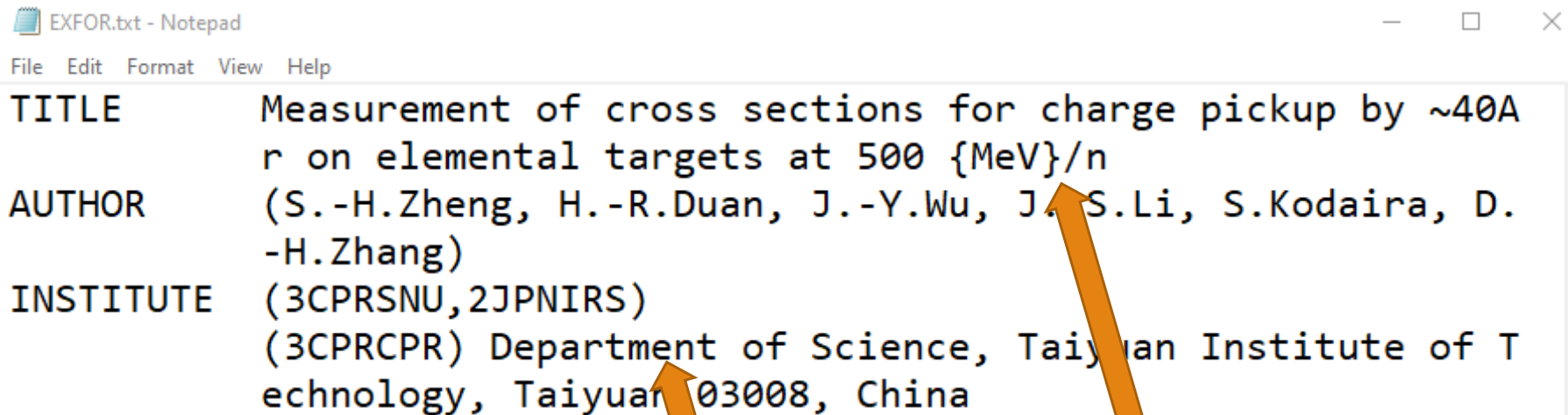
Cross section

CR-39 nuclear track detector

ABSTRACT

The nuclear charge pickup cross sections of ^{40}Ar on polyethylene(CH_2), carbon, aluminum, copper and lead targets at the highest energy of 495 MeV/n were measured using CR-39 nuclear track detector. The cross sections for H were calculated from those measured on C and CH_2 targets. The dependence of charge pickup cross section on target mass was studied, it was found that the nuclear charge pickup cross section exhibits a linear dependence on the target mass.

Output of Example 1





```
EXFOR.txt - Notepad
File Edit Format View Help
TITLE      Measurement of cross sections for charge pickup by ~40A
           r on elemental targets at 500 {MeV}/n
AUTHOR     (S.-H.Zheng, H.-R.Duan, J.-Y.Wu, J.-S.Li, S.Kodaira, D.
           -H.Zhang)
INSTITUTE  (3CPRSNU,2JPNIRS)
           (3CPRCPR) Department of Science, Taiwan Institute of T
           echnology, Taiyuan 03008, China
```

We are working on the refinement of the characters and words in proper format (e.g., TeX symbol, deletion of extra text etc.)

Example 2

PHYSICAL REVIEW C **105**, 054319 (2022)

Evolution of the isoscalar giant monopole resonance in the Ca isotope chain

S. D. Olorunfunmi ^{1,*} R. Neveling,^{2,†} J. Carter,¹ P. von Neumann-Cosel ³ I. T. Usman,¹ P. Adsley,^{1,2,4,5} A. Bahini,¹ L. P. L. Baloyi,¹ J. W. Brümmer,⁴ L. M. Donaldson,^{1,2} H. Jivan,¹ N. Y. Kheswa,¹ K. C. W. Li,⁴ D. J. Marín-Lámbarri,⁶ P. T. Molema,¹ C. S. Moodley,¹ G. G. O'Neill,⁶ P. Papka,^{2,4} L. Pellegrini,^{1,2} V. Pesudo,⁶ E. Sideras-Haddad,¹ F. D. Smit,² G. F. Steyn,² A. A. Avaa,^{1,2} F. Diel,⁷ F. Dunkel,⁷ P. Jones,² and V. Karayonchev⁷

¹*School of Physics, University of the Witwatersrand, Johannesburg 2050, South Africa*

²*iThemba Laboratory for Accelerator Based Sciences, Somerset West 7129, South Africa*

³*Institute für Kernphysik, Technische Universität Darmstadt, D-64289 Darmstadt, Germany*

⁴*Department of Physics, University of Stellenbosch, Matieland 7602, South Africa*

⁵*Institut de Physique Nucléaire d'Orsay, IN2P3-CNRS, Université Paris Sud, Orsay, France*

⁶*Department of Physics, University of the Western Cape, Bellville 7535, South Africa*

⁷*Institute für Kernphysik, Universität zu Köln, D-50937 Köln, Germany*



(Received 30 January 2022; accepted 16 May 2022; published 31 May 2022)

Background: Two recent studies of the evolution of the isoscalar giant monopole resonance (ISGMR) within

Output of Example 2

TITLE Evolution of the isoscalar giant monopole resonance in the Ca isotope chain

AUTHOR (S.Olorunfunmi, R.Neveling, J.Carter, P.Neumann-Cosel, I.Usman, P.Adsley, A.Bahini, L.Baloyi, J.Brümmer, L.Donaldson, H.Jivan, N.Kheswa, K.Li, D.Mar{\'\i}n-L{\'a}mbarri, P.Molema, C.Moodley, G.O{\textquotesingle}Neil, P.Papka, L.Pellegrini, V.Pesudo, E.Sideras-Haddad, F.Smit, G.Steyn, A.Avaa, F.Diel, F.Dunkel, P.Jones, V.Karayonchev)

INSTITUTE (3SAFITH,2GERTHD,3SAFUSF,2FR PAR,3SAFUWC,2GERKLN)
(3SAFSAF) School of Physics, University of the Witwatersrand, Johannesburg 2050, South Africa

Example 3



ELSEVIER



Available online at www.sciencedirect.com

ScienceDirect

Nuclear Physics A 1021 (2022) 122421

**NUCLEAR
PHYSICS** **A**

www.elsevier.com/locate/nuclphysa

Effect of projectile structure on break-up fusion for $^{14}\text{N} + ^{175}\text{Lu}$ system at intermediate energies

Ishfaq Majeed Bhat ^{a,*}, Mohd. Shuaib ^a, M. Shariq Asnain ^a,
Vijay R. Sharma ^b, Abhishek Yadav ^c, Manoj Kumar Sharma ^d,
Pushpendra P. Singh ^e, Devendra P. Singh ^f, Unnati Gupta ^g,
Rudra N. Sahoo ^e, Arshiya Sood ^e, Malika Kaushik ^e, R. Kumar ^h,
B.P. Singh ^a, R. Prasad ^a

^a Nuclear Physics Laboratory, Department of Physics, Aligarh Muslim University, Aligarh-202 002, U.P., India

^b Instituto de Fisica, Universidad Nacional Autonoma de Mexico, Mexico D.F. 01000, Mexico

^c Department of Physics, Faculty of Natural Sciences, Jamia Milia Islamia, New Delhi-110025, India

^d Department of Physics, Shri Varshney College, Aligarh-202 001, U.P., India

^e Department of Physics, Indian Institute of Technology, Ropar, Punjab-140 001, India

^f Department of Physics, University of Petroleum and Energy Studies, Dehradun-248 007, U.K., India

^g Amity Institute of Nuclear Science and Technology, Amity University, Uttar Pradesh, Noida-201313, India

^h NP-Group, Inter University Accelerator Center, Aruna Asaf Ali Marg, New Delhi-110 067, India

Received 23 December 2021; received in revised form 11 February 2022; accepted 14 February 2022


Available online 22 February 2022

Output of Example 3

TITLE Effect of projectile structure on break-up fusion for $^{14}\text{N} + ^{175}\text{Lu}$ system at intermediate energies
AUTHOR (I. Bat, Mohd. Shuaib, M. Asnain, V. Sharma, A. Yadav, M. Sharma, P. Singh, D. Singh, U. Gupta, R. Sahoo, A. Sood, M. Kaur, R. Kumar, B. P. Singh, R. Prasad)
INSTITUTE (3INDMUA, 3MEXUMX, 3INDIIR, 3INDMGA, 3INDNSD)
(3INDIND) Department of Physics, Faculty of Natural Sciences, Jamia Milia Islamia, New Delhi-110025, India
(3INDIND) Department of Physics, Shri Varshney College, India
University of Petr 08 007, U.K., I


We are working on the refinement of the characters and words in proper format (e.g., TeX symbol, deletion of extra text etc.)


Role of the entrance channel in the experimental study of incomplete fusion of ^{13}C with ^{93}Nb

Avinash Agarwal ^{*}, Anuj Kumar Jashwal,[†] Munish Kumar, and S. Prajapati
Department of Physics, Bareilly College, Bareilly (U.P.) 243 005, India


Sunil Dutt, Muntazir Gull, and I. A. Rizvi
Department of Physics, Aligarh Muslim University, Aligarh (U.P.) 202 002, India


Kamal Kumar
Department of Physics, Hindu College, Moradabad (U.P.) 244 001, India

Sabir Ali 
MANUU Polytechnic Darbhanga, Maulana Azad National Urdu University, Hyderabad 500 032, India

Abhishek Yadav 
Department of Physics, Jamia Millia Islamia, New Delhi 110 025, India

R. Kumar
NP-Group, Inter University Accelerator Center, New Delhi 110 067, India

A. K. Chaubey 
Department of Physics, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia

 (Received 9 June 2021; revised 16 January 2022; accepted 25 February 2022; published 10 March 2022)

We report measured cross-section data of the residues produced in the ^{13}C -induced reaction on ^{93}Nb within

Output of Example 4

TITLE Role of the entrance channel in the experimental study
 of incomplete fusion of ^{13}C with ^{93}Nb

AUTHOR (A.Agarwal, A.Jashwal, M.Kumar, S.Prajapati, S.Dutt, M.
 Gull, I.Rizvi, K.Kumar, S.Ali, A.Yadav, R.Kumar, A.Chau
 bey)

INSTITUTE (3INDMUA,3INDNSD)
 (3INDIND) Department of Physics, Bareilly College, Bare
 illy (U.P .) 243 005, India

 (3INDIND) Department of Physics, Hindu College, Morada
 bad (U.P .) 244 001, India


 (3INDIND) MANUU Polytechnic Darbhanga, Maulana Azad Na
 tional Urdu University, Hyderabad 500 032, India

 (3INDIND) Department of Physics, Jamia Millia Islamia,
 New Delhi 110 025, India

Limitations

The code needs extensive refinement and could improve with testing on various research papers and journals.

Few limitations encountered in above four examples:

- Institute names written in language other than English give error sometimes.
 - Could not provide author names on files in which DOI is missing.
 - TRANS Dictionary file has also few short comes.
- 

Institute names in language other than English

- ¹*School of Physics, University of the Witwatersrand, Johannesburg 2050, South Africa*
²*iThemba Laboratory for Accelerator Based Sciences, Somerset West 7129, South Africa*
³*Institute für Kernphysik, Technische Universität Darmstadt, D-64289 Darmstadt, Germany*
⁴*Department of Physics, University of Stellenbosch, Matieland 7602, South Africa*
⁵*Institut de Physique Nucléaire d'Orsay IN2P3-CNRS, Université Paris Sud, Orsay, France*
⁶*Department of Physics, University of the Western Cape, Bellville 7535, South Africa*
⁷*Institute für Kernphysik, Universität zu Köln, D-50937 Köln, Germany*



(Received 30 January 2022; accepted 16 May 2022; published 30 May 2022)

trans - Notepad

File Edit Format View Help

1995 - 2008: Forschungszentrum Karlsruhe (FZK)
2GERKIG (GKSS, Geesthacht)
2GERKIL (Univ. of Kiel, Kiel)
2GERKLN (Universitaet zu Koeln)
2GERKRU (Karlsruhe, Univ.)
Incl. Inst.fuer Struktur der Materie, fuer

Difference in the style of name written in article and dictionary

Institute names in language other than English

The problem could be overcome by

- searching text with initial and last alphabets in the city

OR

- With the input from the compiler

```
def GERMANY(text):  
    import re  
    if re.search(r'\bK\w*?ln\b', text):  
        text=re.sub(r'\bK\w*?ln\b', 'Koeln', text)  
        return str('Koeln'), text  
    return
```

Short comes in TRANS Dictionary file

- Abbreviations in the Institute code are not uniform. For example,
 - University, Universitaet and Univ. are used at different places

2ITYUTV	(Univ.degli Studi di Roma "Tor Verga Compare 2ITYROM
2JAPFE	(Fuji Electric)
2JAPHIR	(Hiroshima University of Hiroshima)
2JAPHOS	(Hosei University, Tokyo)
2JAPUVO	(Hosei University, Tokyo)

2GERBER	(Hahn-Meitner-Inst., Berlin)
2GERBOC	(Ruhr-Universitaet Bochum)
2GERBON	(Univ. of Bonn)
2GERDKZ	(Deutsches Krebsforschungszentrum German Cancer Res. Centre, Heidelberg)
2GERDOR	(Dortmund Univ., F.R.Germany)
2GERDRE	(Dresden, Techn.Univ.)

Abbreviations in the institute code are not uniform. For example,


- University, Universitaet and Univ. are used at different places
- Institute, Inst. etc.

2GERKRU	(Karlsruhe, Univ.)	3000000300523
	Incl. Inst.fuer Struktur der Materie, fuer Angewandte	3000000300524
	Physik, fuer Angew. Kernphysik	3000000300525
2GERLMU	(Ludwig-Maximilians Universitaet Muenchen)	3000000300526
2GERMBG	(Univ. of Marburg)	3000000300527
2GERMNZ	(Johannes Gutenberg-Universitaet Mainz, Mainz)	3000000300528C
2GERMPH	(Max-Planck-Institut fuer Kernphysik, Heidelberg)	3000000300529
	Independent of HEI.	3000000300530
2GERMPM	(Max-Planck-Institut fuer Chemie, Mainz)	3000000300531
2GERMST	(Westfaelische Wilhelms-Universitaet Muenster, Muenster)	3000000300532
	Also known as Universitaet Muenster	3000000300533
2GERMUE	(Muenchen, Techn.Univ.)	30000003005340
	Obsolete. Use 2GERMUN.	30000003005350
2GERMUN	(Technische Universitaet Muenchen)	3000000300536
	Formerly Technische Hochschule Muenchen	3000000300537
	Including Forschungs-Neutronenquelle	3000000300538
	Heinz-Maier-Leibnitz, Garching	


Institute codes given in
comments could also be
properly coded

TRANS Dictionary

Above limitations were encountered while running the code for the examples. There could be more in future. Moreover

- Trans file could work better if modified in form of CSV file with the columns consisting of city names.
 - CSV file could also be used to write the full name of the institute rather than fitting it in 55 columns.
- 

Future work that could be done with collaborations

- In case there is confusion between the Institute codes the compiler could be asked to choose the best option from the selected institutes.
 - The output could be saved in the database for future reference to make the code self learning from the inputs received.
- 

Future work that could be done with collaborations

The code could also be modified

- To add the journal reference in the entry
- To search for the FACILITY given in the research paper.
- To search for a code under other keywords (e.g., DETECTOR, METHOD)

Acknowledgement

I would like to thank Dr. Naohiko for providing the idea and the fruitful discussions.

Thank you

for your kind attention