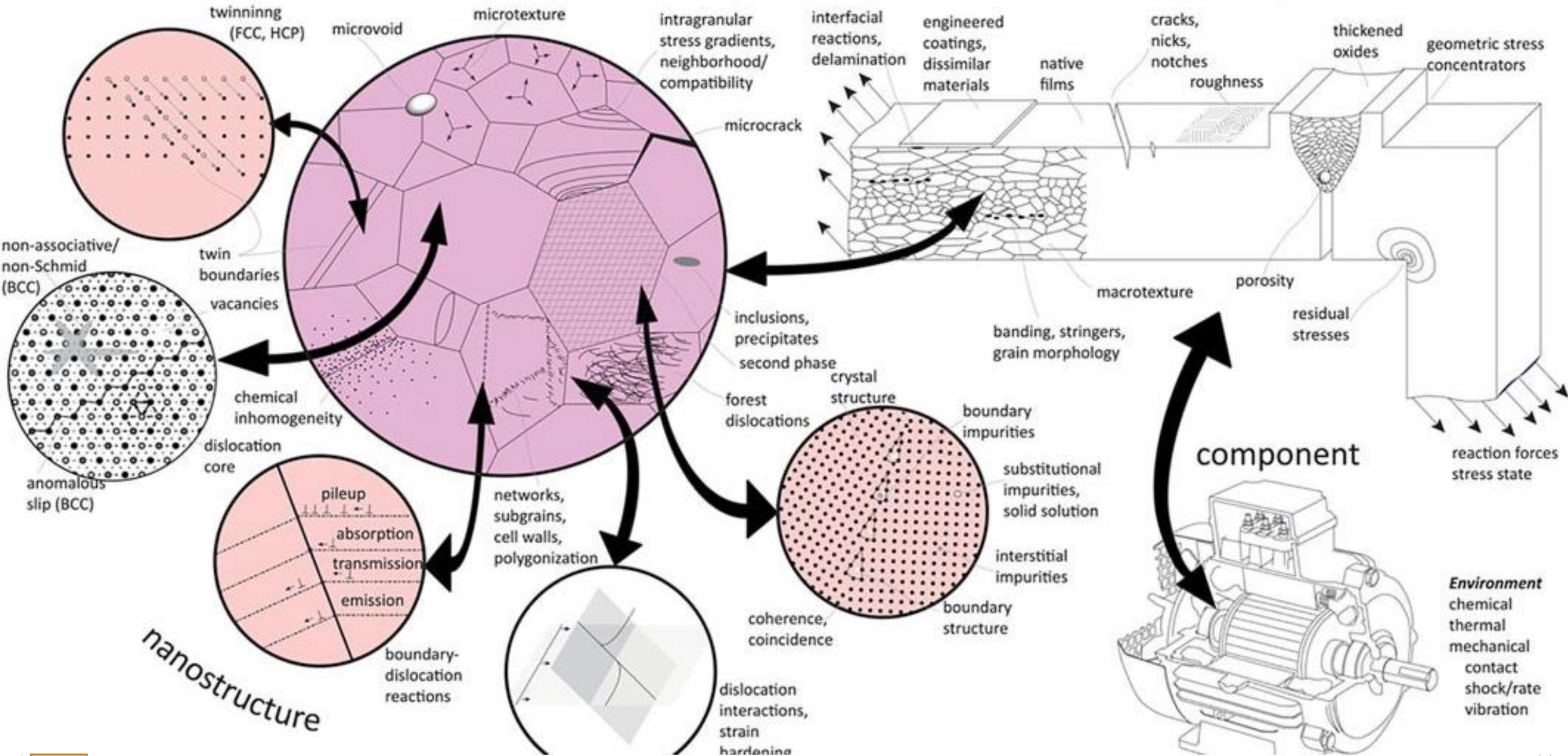


Парсеры **.CIF** и **.OUT** файлов

Подготовил студент группы КЭ-222 Рассказов Сергей Михайлович
Научный руководитель: к.т.н. Кафтанников Игорь Леопольдович



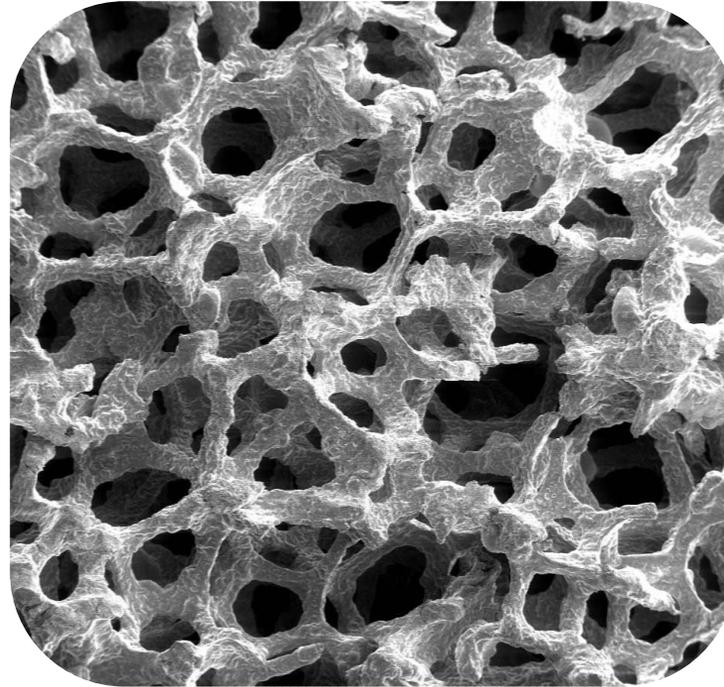
Digital Twins for Materials

Технологический
институт Джорджии;
Surya R. Kalidindi, Michael
Buzzy, Brad L. Boyce,
Remi Dingreville 16.03.22г.



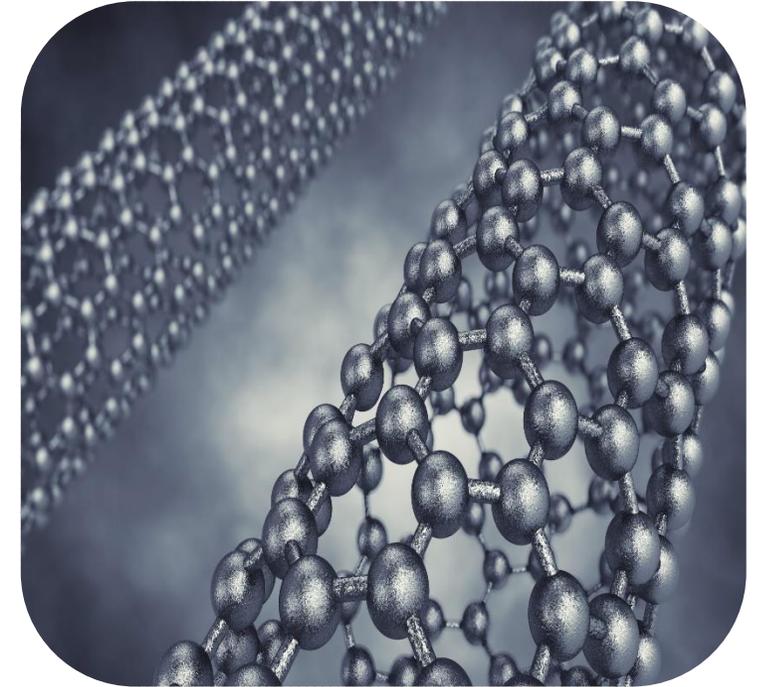
МАКРОУРОВЕНЬ

Физические объекты в метриках метровых,
сантиметровых единиц



МИКРОУРОВЕНЬ

Сплавы, твердые растворы, органические
соединения, агрегации молекул и т.п.



НАНОУРОВЕНЬ

В основном, структуры с нанометрикой,
молекулы

АТОМНЫЙ УРОВЕНЬ

Допустимо опираться на эти три уровня, если рассматриваются довольно большие объекты, так в этой статье речь идёт о газотурбинных двигателях. В случае, если речь идёт о цифровых двойниках материалов, то необходимо опускаться на более низкий уровень – атомный. При такой детализации определяются и исследуются химические связи, и силы этих связей. Свойства взаимного расположения атомов

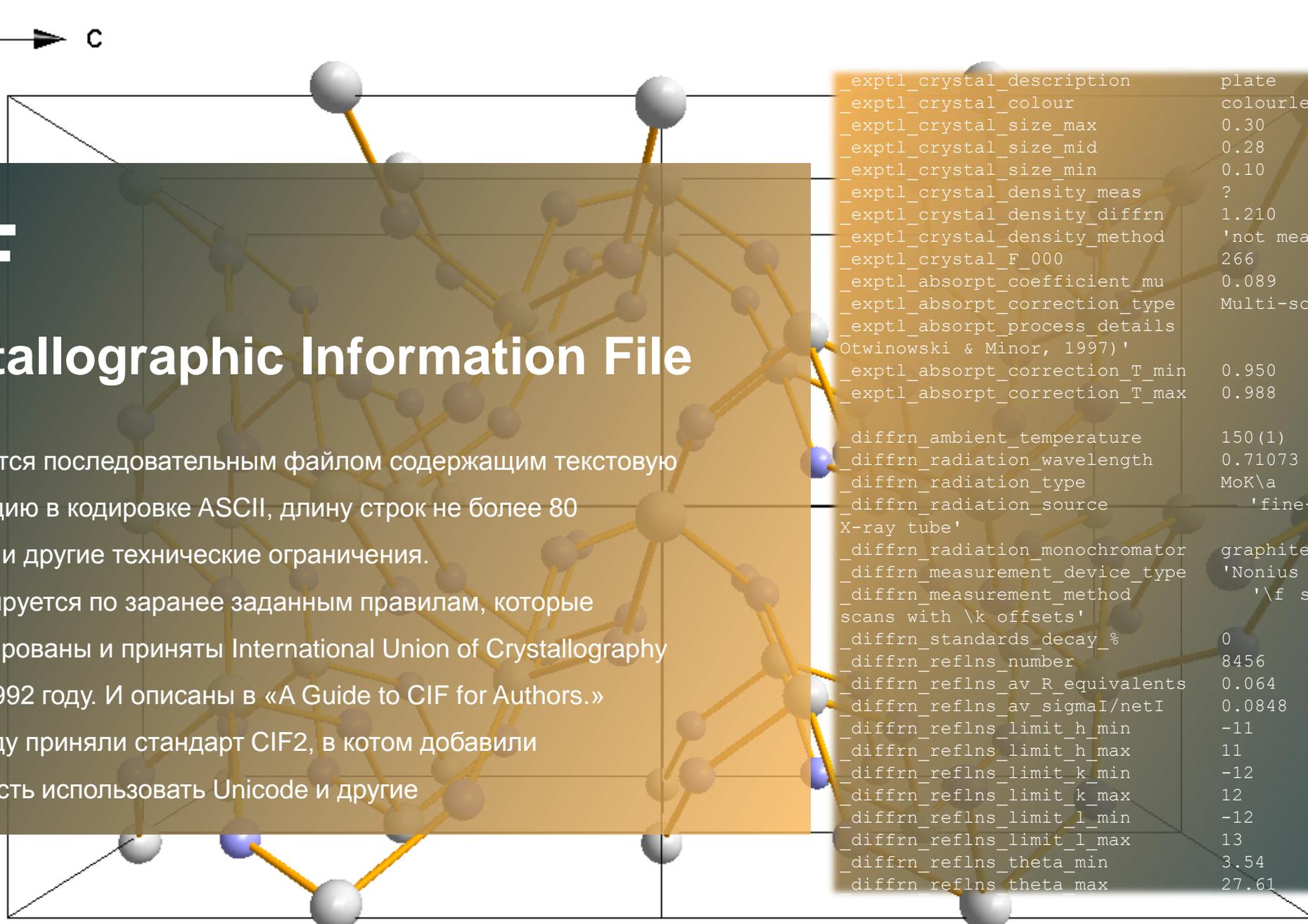
CIF

Crystallographic Information File

CIF является последовательным файлом содержащим текстовую информацию в кодировке ASCII, длину строк не более 80 символов и другие технические ограничения.

Он формируется по заранее заданным правилам, которые сформулированы и приняты International Union of Crystallography (IUCr) в 1992 году. И описаны в «A Guide to CIF for Authors.»

В 2014 году приняли стандарт CIF2, в котом добавили возможность использовать Unicode и другие



```
_exptl_crystal_description      plate
_exptl_crystal_colour          colourless
_exptl_crystal_size_max        0.30
_exptl_crystal_size_mid        0.28
_exptl_crystal_size_min        0.10
_exptl_crystal_density_meas    ?
_exptl_crystal_density_diffrn  1.210
_exptl_crystal_density_method  'not measured'
_exptl_crystal_F_000           266
_exptl_absorpt_coefficient_mu   0.089
_exptl_absorpt_correction_type  Multi-scan
_exptl_absorpt_process_details '(DENZO-SMN;
Otwinowski & Minor, 1997)'
_exptl_absorpt_correction_T_min 0.950
_exptl_absorpt_correction_T_max 0.988

_diffraction_ambient_temperature 150(1)
_diffraction_radiation_wavelength 0.71073
_diffraction_radiation_type       MoK\alpha
_diffraction_radiation_source     'fine-focus sealed
X-ray tube'
_diffraction_radiation_monochromator graphite
_diffraction_measurement_device_type 'Nonius KappaCCD'
_diffraction_measurement_method     '\f scans, and \w
scans with \k offsets'
_diffraction_standards_decay_percent 0
_diffraction_reflns_number         8456
_diffraction_reflns_av_R_equivalents 0.064
_diffraction_reflns_av_sigmaI/netI 0.0848
_diffraction_reflns_limit_h_min    -11
_diffraction_reflns_limit_h_max    11
_diffraction_reflns_limit_k_min    -12
_diffraction_reflns_limit_k_max    12
_diffraction_reflns_limit_l_min    -12
_diffraction_reflns_limit_l_max    13
_diffraction_reflns_theta_min      3.54
_diffraction_reflns_theta_max      27.61
```





ATOMS IN THE ASYMMETRIC UNIT 5 - ATOMS IN THE UNIT CELL: 16

ATOM		X/A	Y/B	Z/C
1 T	6 C	0.000000000000E+00	5.000000000000E-01	3.264
2 F	6 C	-5.000000000000E-01	0.000000000000E+00	
3 T	8 O	0.000000000000E+00	5.000000000000E-01	
4 F	8 O	-5.000000000000E-01	0.000000000000E+00	
5 T	7 N	1.460055586849E-01	-3.539944413151E-01	
6 F	7 N	-1.460055586849E-01	3.539944413151E-01	
7 F	7 N	-3.539944413151E-01	-1.460055586849E-01	
8 F	7 N	3.539944413151E-01	1.460055586849E-01	
9 T	1 H	2.588485243130E-01	-2.411514756870E-01	
10 F	1 H	-2.588485243130E-01	2.411514756870E-01	
11 F	1 H	-2.411514756870E-01	-2.588485243130E-01	
12 F	1 H	2.411514756870E-01	2.588485243130E-01	
13 T	1 H	1.436313742521E-01	-3.563686257479E-01	
14 F	1 H	-1.436313742521E-01	3.563686257479E-01	
15 F	1 H	-3.563686257479E-01	-1.436313742521E-01	
16 F	1 H	3.563686257479E-01	1.436313742521E-01	

T = ATOM BELONGING TO THE ASYMMETRIC UNIT
INFORMATION **** fort.34 **** GEOMETRY OUTPUT FILE

DIRECT LATTICE VECTORS CARTESIAN COMPONENTS (ANGSTROM)			
X	Y	Z	
0.556500000000E+01	0.000000000000E+00	0.000000000000E+00	
0.000000000000E+00	0.556500000000E+01	0.000000000000E+00	
0.000000000000E+00	0.000000000000E+00	0.468400000000E+01	

CARTESIAN COORDINATES - PRIMITIVE CELL

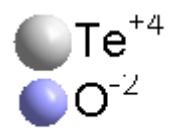
* ATOM	X (ANGSTROM)	Y (ANGSTROM)	Z (ANGSTROM)
1	6 C	0.000000000000E+00	2.782500000000E+00
2	6 C	2.782500000000E+00	0.000000000000E+00
3	8 O	0.000000000000E+00	2.782500000000E+00

OUT

Хранение результатов расчетов

Является результатом работы программных пакетах по расчету электронной структуры материалов. Формат стал популярным из-за постоянного развития в области химии и материалов.

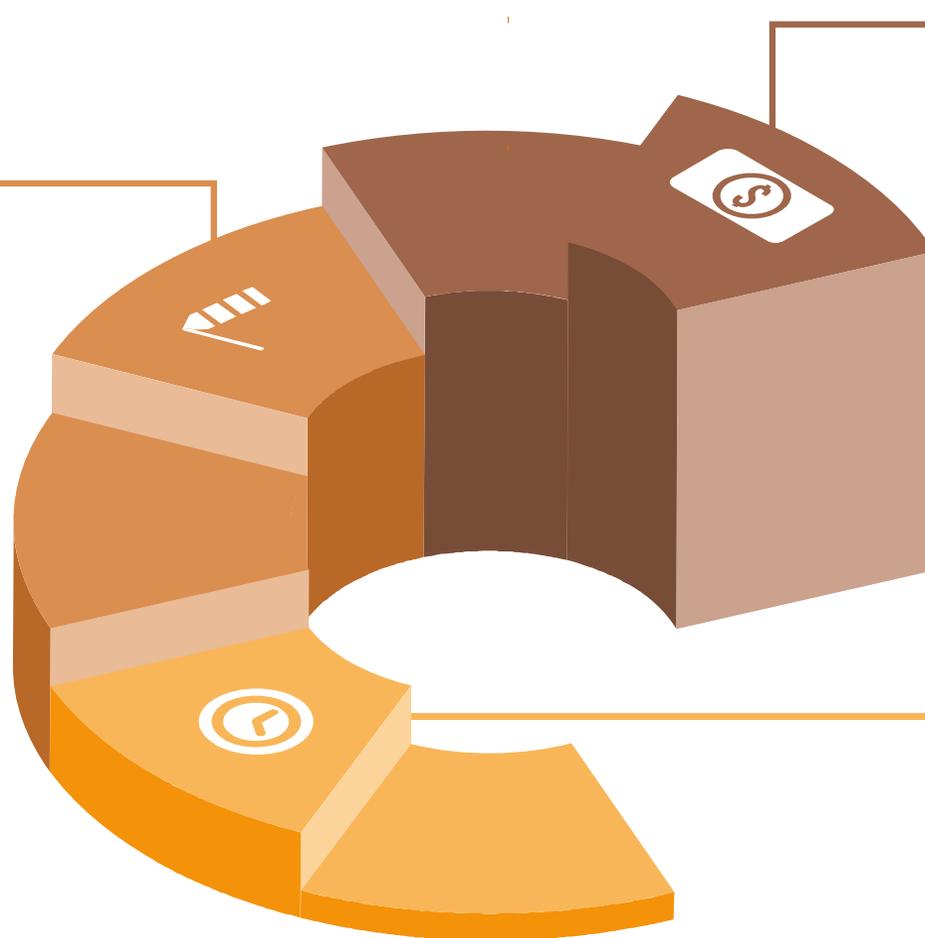
Старые форматы уже не могут покрывать все требования современных исследований. Основным отличием от CIF файла является отсутствие правил по формированию данных, каждая программа делает это по своему, хотя общие черты конечно прослеживаются.



Парсинг

Парсинг

Парсинг – это процесс автоматизированного сбора и структурирование информации из источника при помощи программы или сервиса, для дальнейшей работы с ней, как с отдельными объектами.



Результат

Результатом парсинга является структурированные данные, которые были извлечены из исходного источника информации. Эти данные могут быть представлены в различных форматах, в зависимости от того, какой тип информации вы извлекаете и какой инструмент для парсинга используете.

Источник данных

Источником парсинга может быть любой источник данных, содержащий информацию, которую вы хотите извлечь. Веб-страницы, базы данных, файлы CSV или JSON, XML-документы, файлы логов и др.

Группы



default_group

01

Данные о дате создания документа



submission_details

02

Данные об авторах



processing_summary

03

Данные о журнале, в котором опубликовано исследование



title_and_author

04

Понятное название и аннотация



text

05

Данные к рисункам



chemical_data

06

Брутто формула, данные о симметрии, параметрах кристалла и

т.д.



refinement_data

07

Уточняющие данные, схема весов, параметры матриц, коэффициенты



atomic_coordinates_and_displacement_parameters

08

Данные об атомах, их расположение и т.д.



molecular_geometry

09

Дополнительная информация, не входящая в основной перечень



Unknown_group

10

Парсинг CIF

	Что ищем	Регулярное выражение
Пары «Ключ-значение»	<pre>_exptl_crystal_size_min 0.10 _exptl_crystal_density_meas ? _exptl_crystal_density_diffn 1.210 _exptl_crystal_density_method 'not measured' _exptl_crystal_F_000 266</pre>	<code>re.match('(_\w+)[\s\t]+(.+)')</code>
Координаты атомов	<pre>N4 0.027(3) 0.031(3) 0.043(3) 0.017(3) 0.014(3) 0.006(2) C5 0.046(4) 0.036(4) 0.047(4) 0.019(3) 0.023(4) 0.016(3) C51 0.067(5) 0.057(5) 0.036(4) 0.023(4) 0.024(4) 0.027(4)</pre>	<code>re.match('^(\w+)\s(\w+)\s([-]*[0-9].[0-9]+)([*\d+])*s([-]*[0-9].[0-9]+)([*\d+])*s([-]*[0-9].[0-9]+)([*\d+])*s*\w+'</code>

АЛГОРИТМ



Источник

Загрузка CIF файла в приложение (сервис)



Поиск ключей

Распознавание пар «Ключ-значение» и формирование объекта



Группировка

Определение группы для пары «Ключ-значение»



Поиск координат

Распознавание строк содержащих координаты атомов



Результат

JSON со всеми распознанными данными разделенные в группы

Следующий файл

РЕЗУЛЬТАТ



Для примера был проведен эксперимент, который показал превосходство разделенной информации по сравнению с последовательной. Данные хранились в локально расположенной базе данных PostgreSQL, на твердотельном накопителе. При работе на обычном жестком диске предполагается большее преимущество. Также при большем количестве данных в базе разделение будет давать всё больший и больший эффект.

Поиск осуществлялся в группе «Chemical Data», по брутто формуле: «C34 H22 N4 O1 S1».

Количество элементов в базе (шт.)	Время поиска в полных файлах (сек)	Время поиска в разделенных файлах	Во сколько раз быстрее
433	0.011000	0.001000	11,0
4753	0.101000	0.004430	22,8

Парсинг OUT

	Что ищем	Принцип поиска
Ключевые слова	urea	<code>self.content.index('urea')</code>
Точные координаты от ключевого слова	CRYSTAL 113 5.565 4.684 5 6 0.000000000 5.000000000 3.255838019 8 0.000000000 5.000000000 -4.028106074 7 1.462928216 -3.537071783 1.761188478 1 2.592098873 -2.407901126 2.809651119 1 1.444918067 -3.555081932 -4.036342956 OPTGEOM ATOMONLY END	<code>object_of_research = self.content[header_start+1]</code> <code>spacegroup = self.content[header_start+3]</code> <code>params_cell_crystal = self.content[header_start+4]</code>
Поиск таблиц	FINAL OPTIMIZED GEOMETRY 1 T 6 C 0.000000000 5.000000000 3.264572093 2 F 6 C -5.000000000 0.000000000 -3.264572093 3 T 8 O 0.000000000 5.000000000 -4.022576805 4 F 8 O -5.000000000 0.000000000 4.022576805 5 T 7 N 1.460055586 -3.539944413 1.775143436	<code>for i in range(self.find_next_end(start_variables), self.content.__len__()):</code> <code>if re.match('\d+', self.content[i]):</code> <code>final_finish = i+1</code>

Типы OUT файлов



OPTimization

Много итерационный расчёт при котором изменяется положение атомов друг относительно друга с целью выявления самого низкого значения энергии



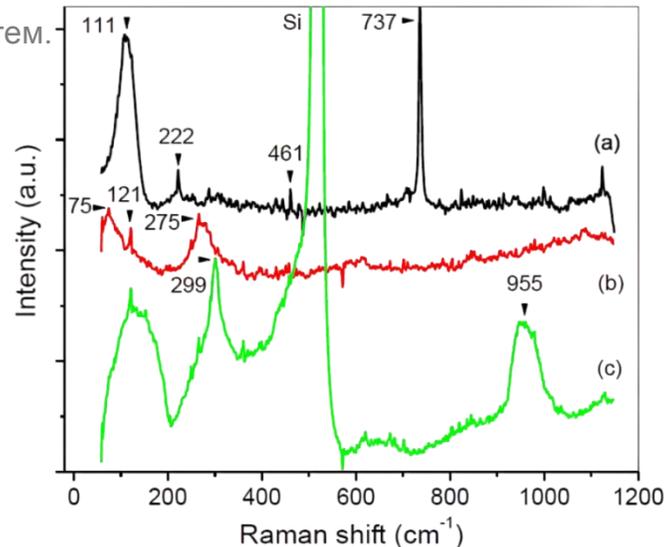
Other

Другие расчеты, которые занимают около 5% от всех расчётов лаборатории



RAMAN спектр

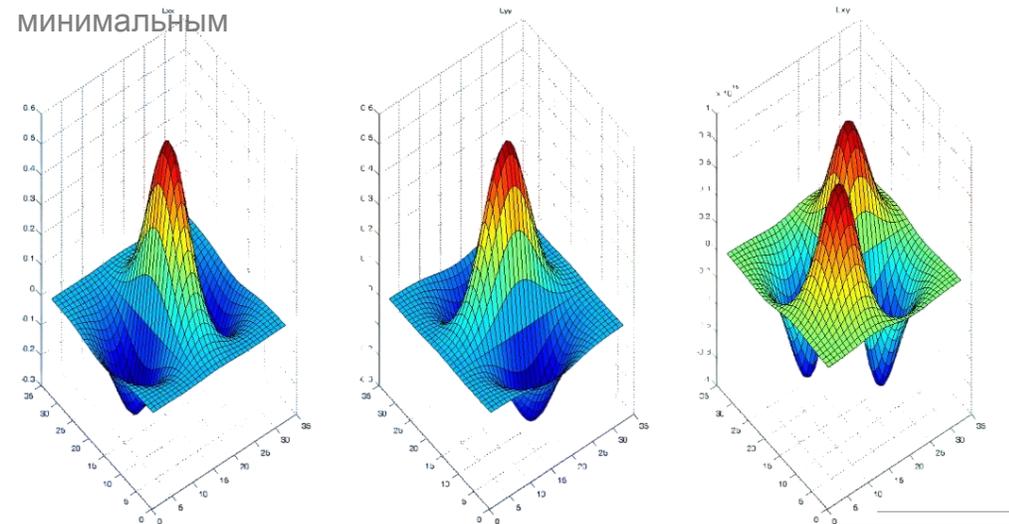
Определение колебательных мод молекул и вибрационных мод в твёрдых телах, который также служит для определения вращательных и других низкочастотных мод систем.



HESSian матрица

По своей сути данный тип расчетов служит для подтверждения данных оптимизации. На сколько, найденное значение в самом деле является

МИНИМАЛЬНЫМ



A composite image featuring laboratory glassware (Erlenmeyer flasks, test tubes) on the left and a microscope on the right, all set against a light blue background with a white diagonal line.

Работа описанная в докладе осуществляется в рамках гранта

«Приоритет 2030, подпроект "Цифровой двойник материалов"»

Лабораторией Многомасштабного моделирования многокомпонентных функциональных материалов

Под руководством доктора химических наук, доцента, Барташевич Екатерины

Владимировны

Спасибо за внимание!