AUTOMATED KOS-BASED SUBJECT INDEXING IN INIS

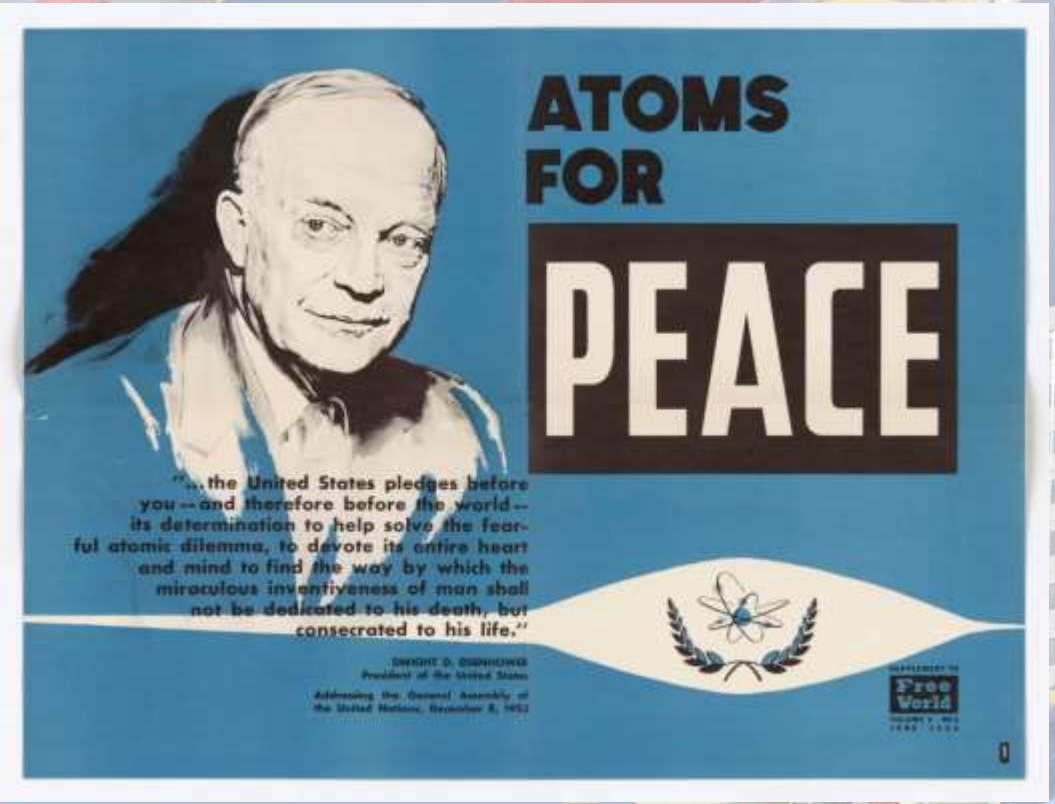
Zaven Hakopov

Dmitry Mironov

Dobrica Savic

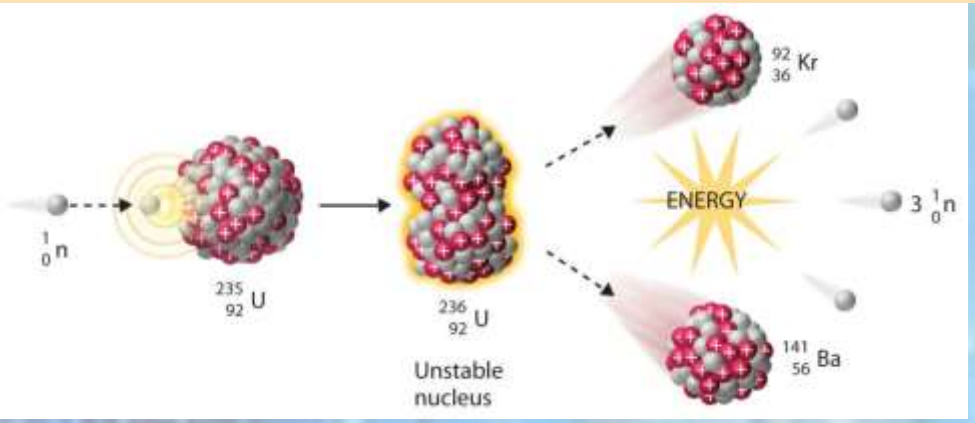
Yulia Svetashova

INTERNATIONAL ATOMIC ENERGY AGENCY (IAEA)

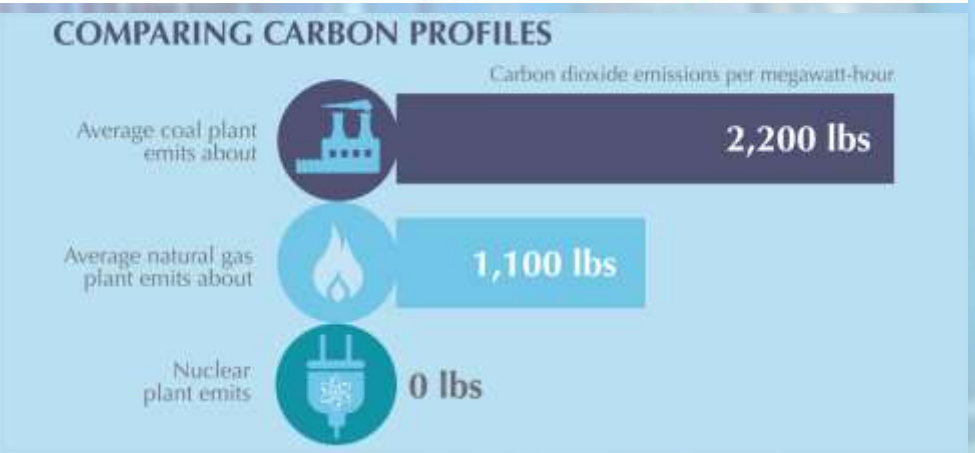


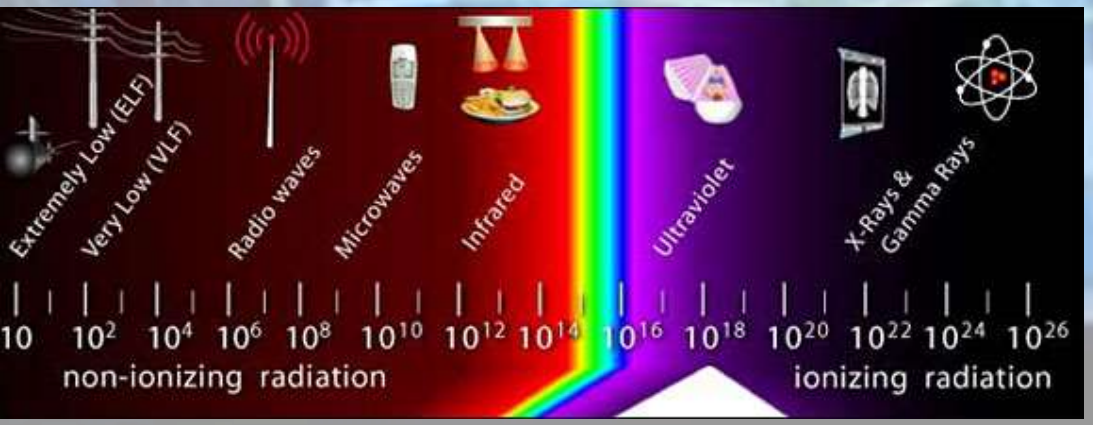
IAEA

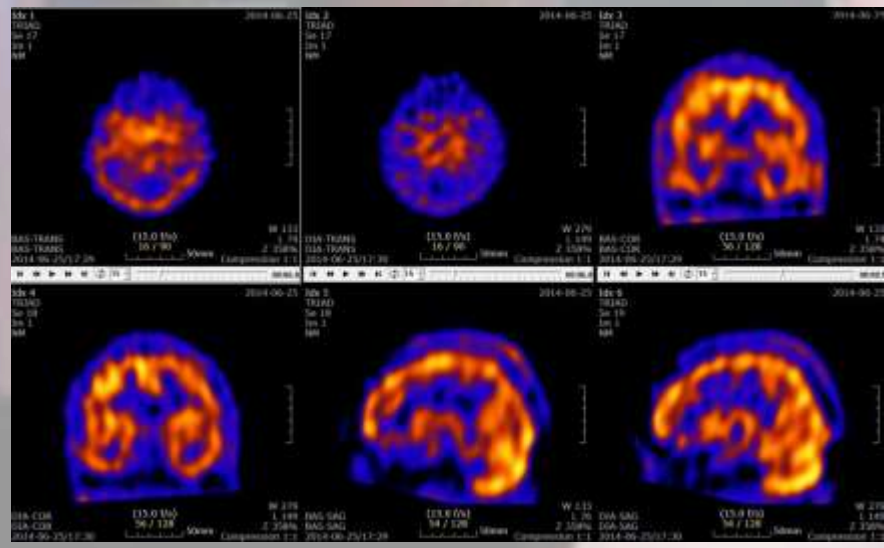
INTERNATIONAL NUCLEAR INFORMATION SYSTEM (INIS)



IAEA | International Nuclear Information System INIS







INTERNATIONAL NUCLEAR INFORMATION SYSTEM (INIS)

155 INIS members

131 Member States

24 international organizations

4 objectives

Acquire, process, preserve literature on peaceful uses of nuclear science and technology

Provide free and open access to nuclear information

Develop and maintain a nuclear knowledge organization system

Assist IAEA Member States in building their scientific information management capacities

INIS repository

4.1 million bibliographic records
104 000 in 2017

540 000 full-text documents
8000 in 2017

2.9 million web page views

1.6 million unique searches

1 million unique visitors

INIS THESAURUS

INIS thesaurus

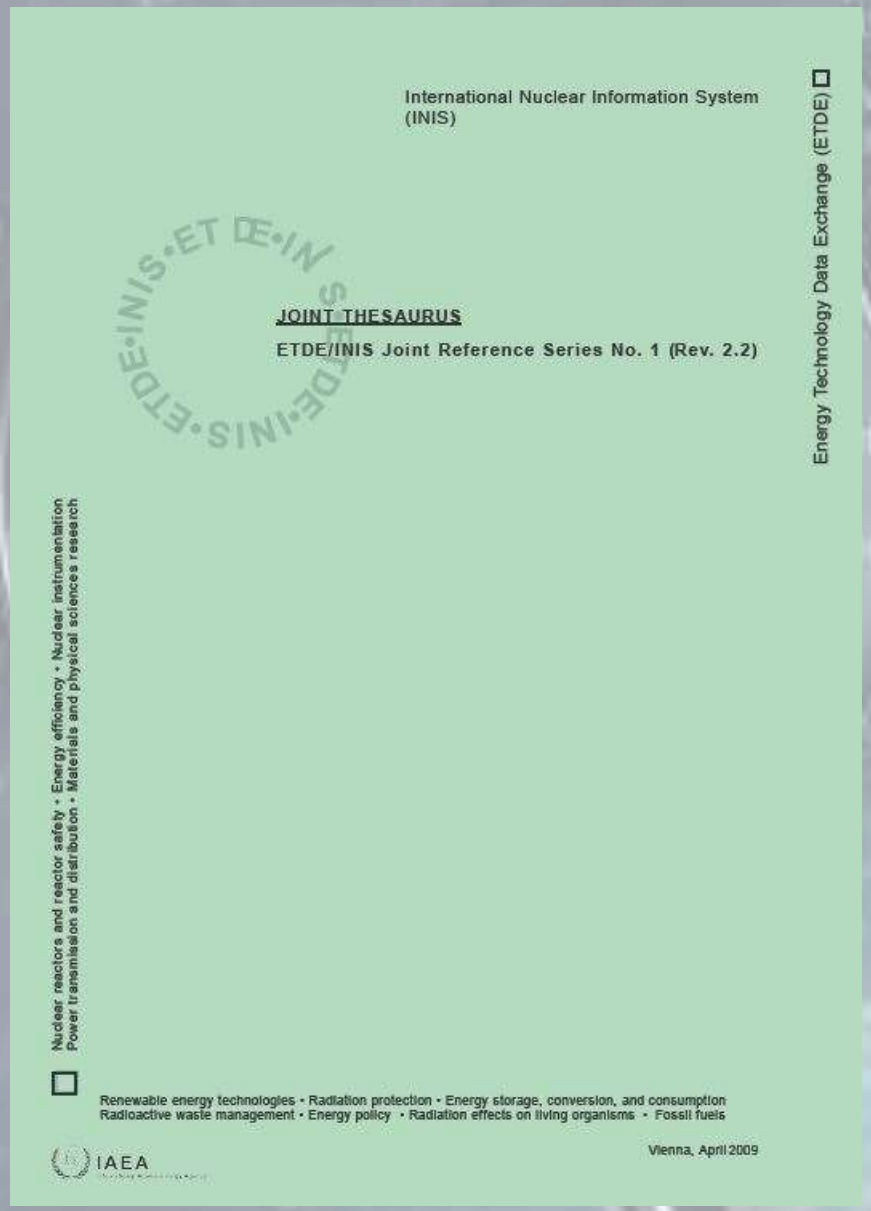
A knowledge organization system

8 languages

Arabic, Chinese, English, French, German, Japanese, Russian, Spanish

31 158 descriptors

150 in 2017



International Nuclear Information System (INIS)

JOINT THESAURUS

ETDE/INIS Joint Reference Series No. 1 (Rev. 2.2)

Energy Technology Data Exchange (ETDE)

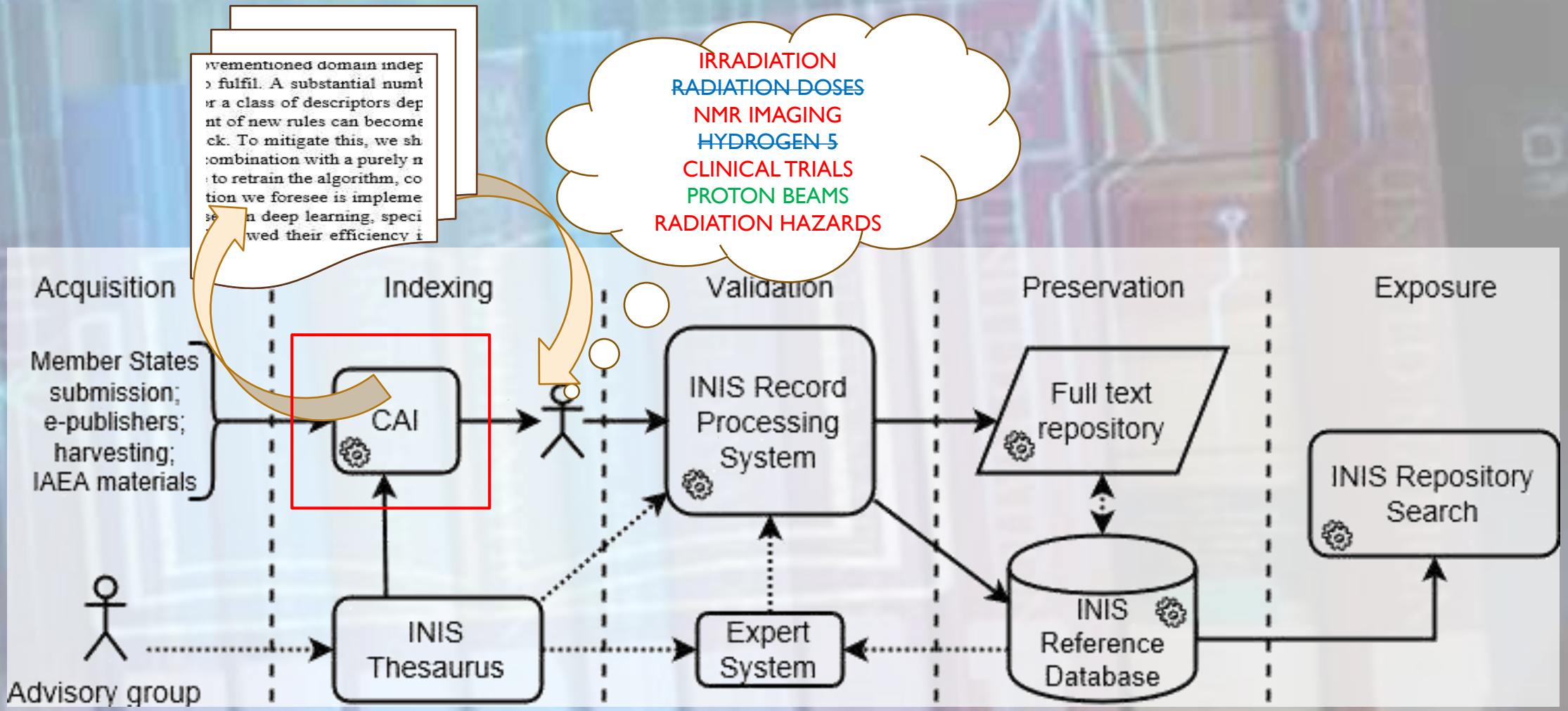
Nuclear reactors and reactor safety • Energy efficiency • Nuclear instrumentation
Power transmission and distribution • Materials and physical sciences research

Renewable energy technologies • Radiation protection • Energy storage, conversion, and consumption
Radioactive waste management • Energy policy • Radiation effects on living organisms • Fossil fuels

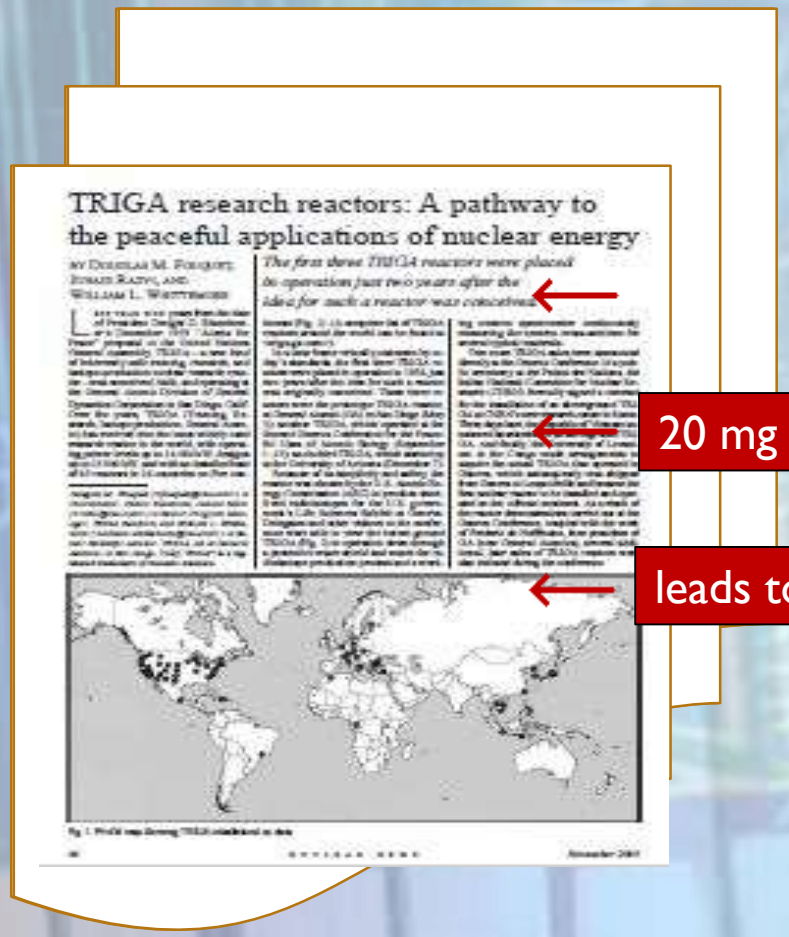
IAEA

Vienna, April 2009

INIS WORKFLOW



CHALLENGES



20 mg

leads to a melting

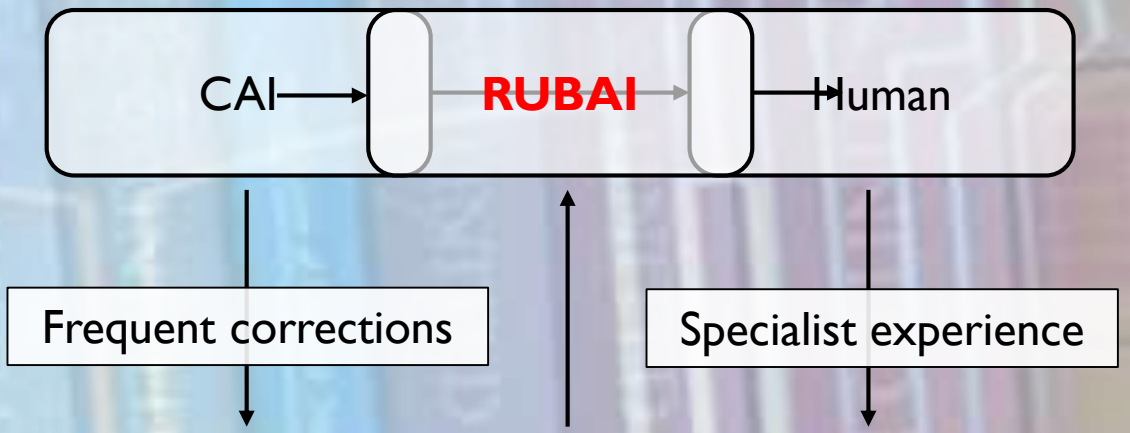
Resulting subject analysis often contains

- Too broad descriptors
- Misleading suggestions
- Incorrect interpretations

45 years of experience

- 4 millions indexed records
- Expert system
- INIS Thesaurus

TWO-PASS INDEXING, TIER I



840 unique rules

- Add
- Replace
- Remove
- Keep

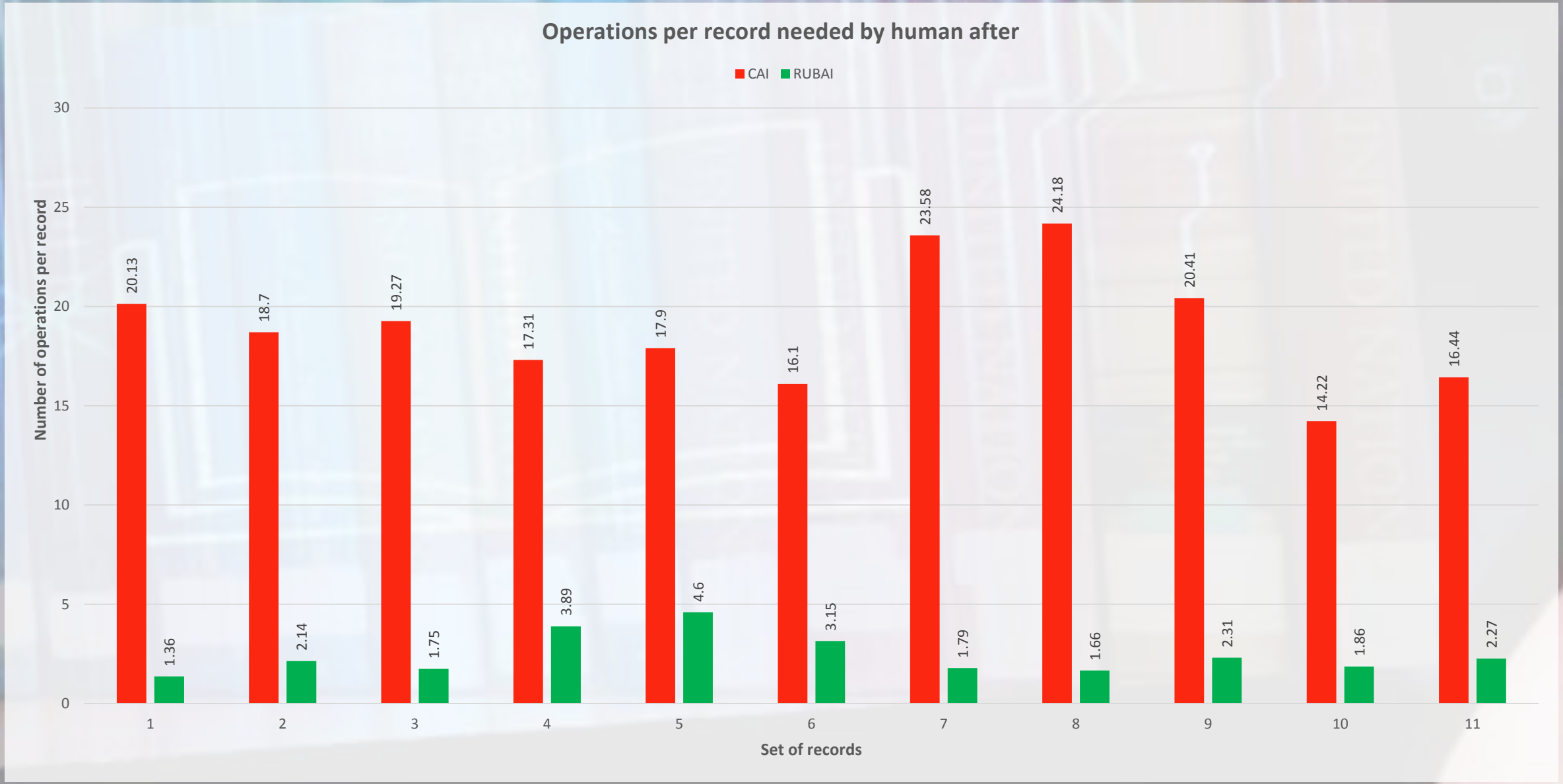
KOS integration

- Thesaurus
- INIS subjects
- Expert knowledge
- Quality control

RUBAI: RESULTS

ID	Number of records	Descriptors								Records where more than one descriptor were				Operations per record performed by human after…	
		Total after…			added by human after…			removed by human after…		added after…		removed after…			
		Average													
		Descriptors											Operations per record performed by human after…		
		Total after…			added by human after…			removed by human after…							
		CAI	RUBAI	Human	CAI	RUBAI		CAI	RUBAI				CAI	RUBAI	
		2280	955	864	114	63		1530	155				18.93	2.43	
1															1.36
2															2.14
3															1.75
4															3.89
5															4.6
6															3.15
7	97	3030	1111	971	114	17	2173	157	31	4	97	39	23.58	1.79
8	62	2019	745	658	69	8	1430	95	15	1	62	23	24.18	1.66
9	93	1799	782	716	99	50	1182	116	22	10	92	35	20.41	2.31
10	98	1912	802	760	121	70	1273	112	36	14	97	30	14.22	1.86
11	77	1712	753	788	171	105	1095	70	53	30	77	21	16.44	2.27

RUBAI: RESULTS

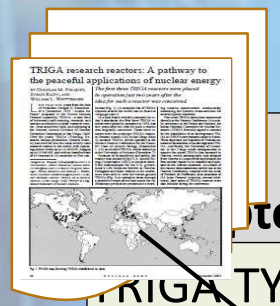


ARE WE DOING WELL?

TRIGA TYPE REACTORS
REACTORS
NEUTRONS
NUCLEAR FUELS
MAGNESIUM 20
REACTOR DESIGN
RESEARCH REACTORS
LEAD
FISSION PRODUCTS
NEUTRON BEAMS
RADIOGRAPHY
MELANOMAS

TRIGA TYPE REACTORS
NUCLEAR FUELS
REACTOR DESIGN
RESEARCH REACTORS
FISSION PRODUCTS
NEUTRON BEAMS
INDUSTRIAL RADIOGRAPHY
~~MELANOMAS~~
NEUTRON CAPTURE THERAPY
ISOTOPE PRODUCTION
TRIGA-2-MUSASHI REACTOR

FROM A SINGLE DECISION TO A DATASET



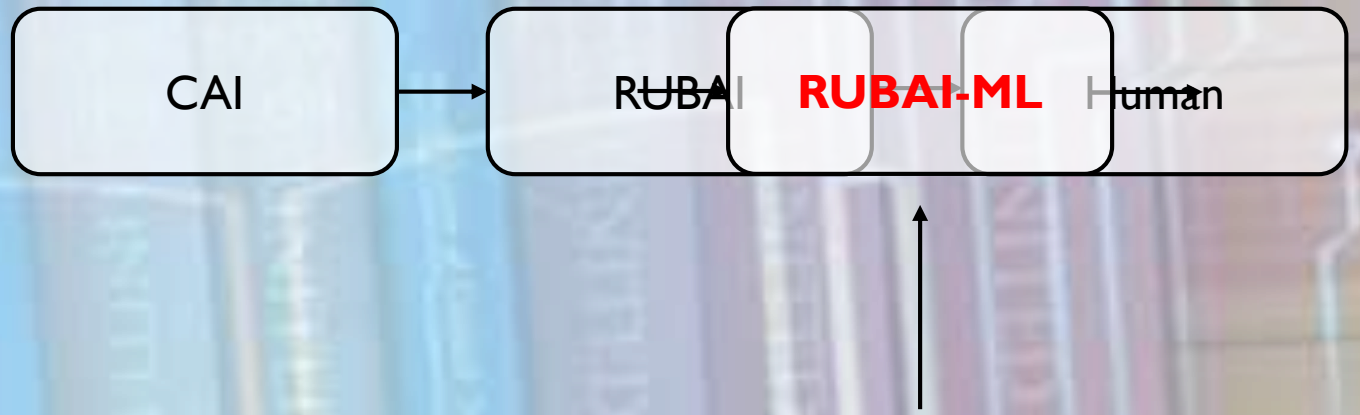
+

tor	RUBAI Operation	Human Action	Human Decision
RIGA TYPE REACTORS	Keep	Keep	Approve
REACTORS	Remove		Approve
RESEARCH REACTORS	Keep	Keep	Approve
LEAD	Remove		Approve
RADIOGRAPHY	Add	Reuse	Reject
M			
N			
IS			
IN			

DATASET

Features					Label
Category	CMV	Weight	Title Match	Operation	Decision
60	-0.0211	0.2474	0	2: remove	1: approve
36	0.5676	0.4759	1	3: keep	0: reject
37	2.5355	0.7849	0	1: add	1: approve

TWO-PASS INDEXING, TIER 2



840 uni

A

R

R

K

DATA analysis

Machine Learning

KOS

Modeling human cognition

DATA set

5,600 documents

84% accuracy

37,000 unique decisions

RUBAI-ML RESULTS: BRINGING THE OUTCOME…

Model evaluation results				
Accuracy	Precision	Recall	Specificity	F_1
0.837	0.838	0.793	0.873	0.815

ID	Number of records	Total after…			Total descriptors after…				man after…		Operations per record performed by human after…	
		CAI	RUBAI	RUBAI-ML	CAI	RUBAI	RUBAI-ML	HUMAN	I	RUBAI-ML	CAI	RUBAI
					2519	1032	1040	1094				
1	120	2519	1032	1040	2104	859	862	936	200	202	*3.85*	*3.83*
2	111	2104	859	862	2217	916	915	960	155	157	*3.49*	*3.49*
3	109	2217	916	915	474	169	172	198	172	175	*3.56*	*3.62*
4	22	474	169	172	**1829**	**744**	**747**	**797**	24	26	*3.5*	*3.54*

… CLOSER TO THE HUMAN CHOICES

RUBAI-ML: RESULTS

Identifies missing descriptors in case of:

- Specific semantic relations
- Narrower descriptors
- Broader document scope

CONCLUSIONS

840

KOS

DATA

DATA set

- 5,600 documents
- 84% accuracy
- 37,000 unique decisions

Work experience
Expert knowledge

+

KOS integration

+

Data analysis
Decision simulation

OUTCOMES AND FUTURE WORK

Future work:

- Extend the coverage and consistency of rules
- Improve ML validation component
- **OR:** replace rules and ML validation with a single ML solution
- **DREAM:** completely data-driven classification algorithm based on (convolutional) neural networks