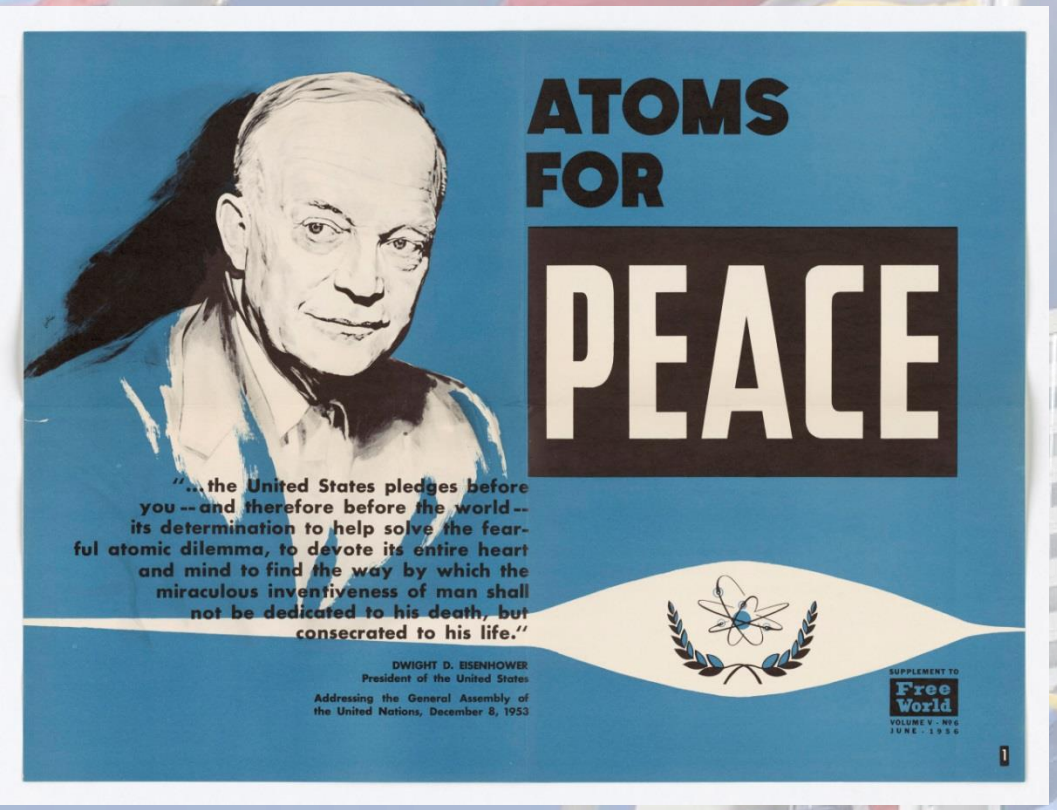
AUTOMATED KOS-BASED SUBJECT INDEXING IN INIS

Dmitry Mironov

Nuclear Information Section

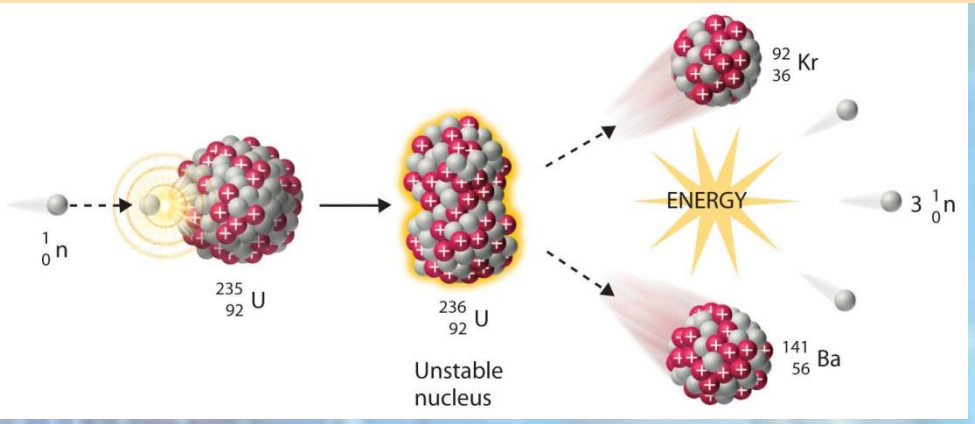
International Atomic Energy Agency (IAEA)

INTERNATIONAL ATOMIC ENERGY AGENCY (IAEA)

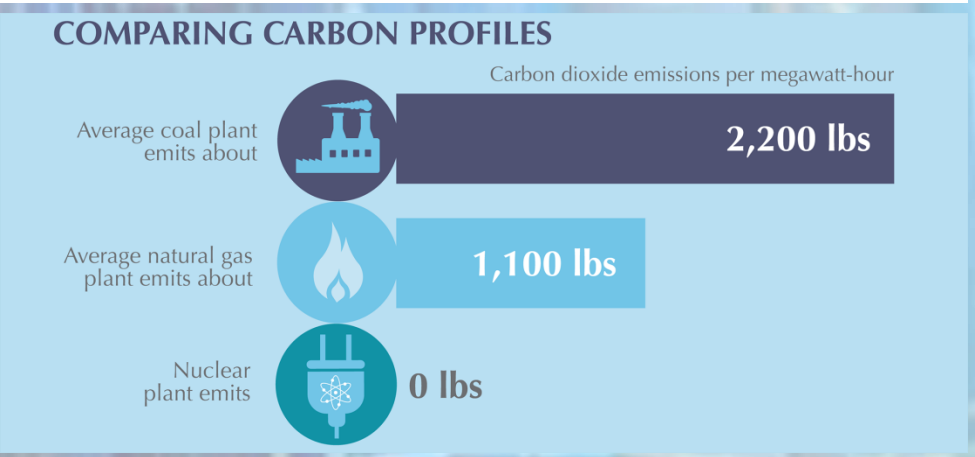


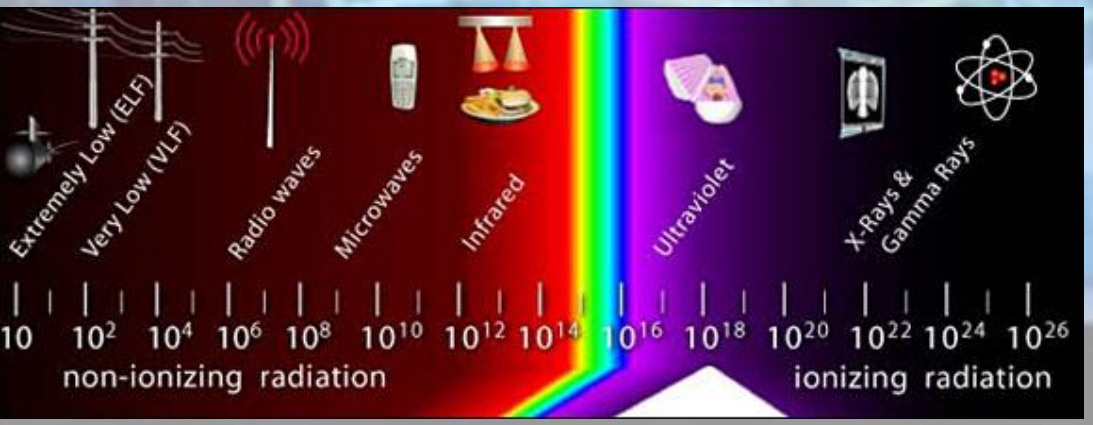
IAEA

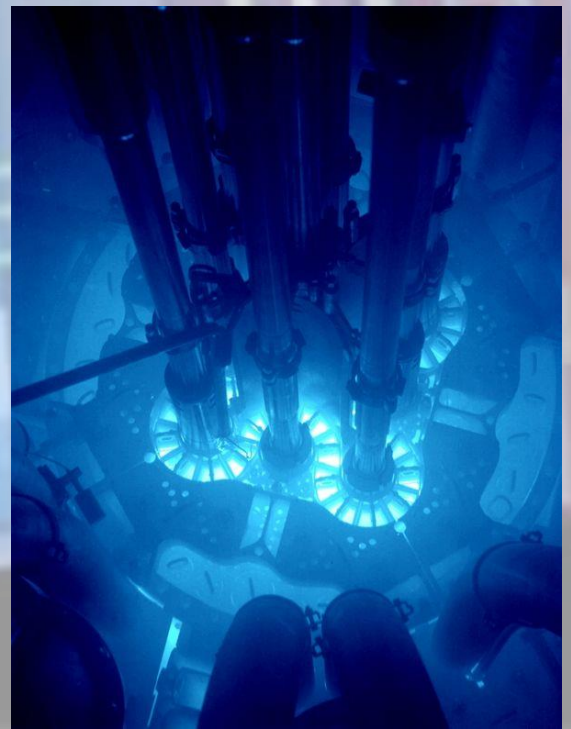
INTERNATIONAL NUCLEAR INFORMATION SYSTEM (INIS)

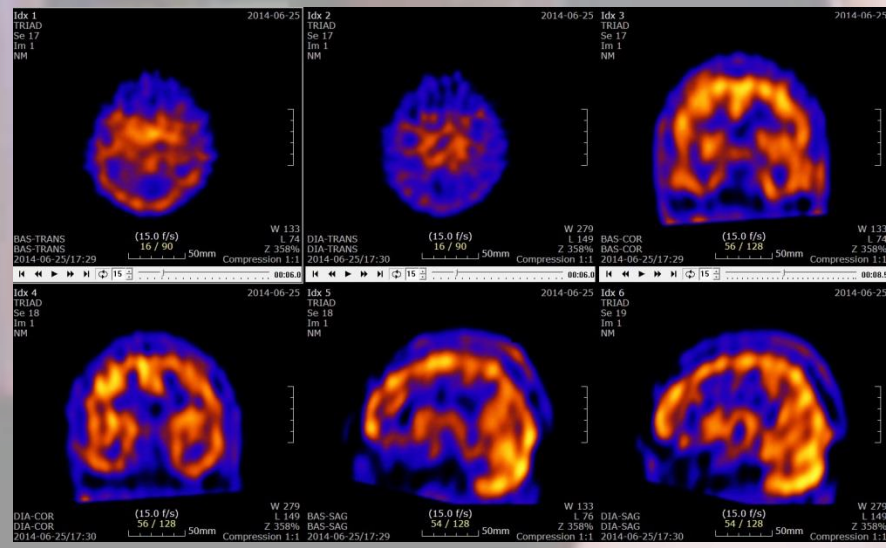


IAEA | International Nuclear Information System INIS





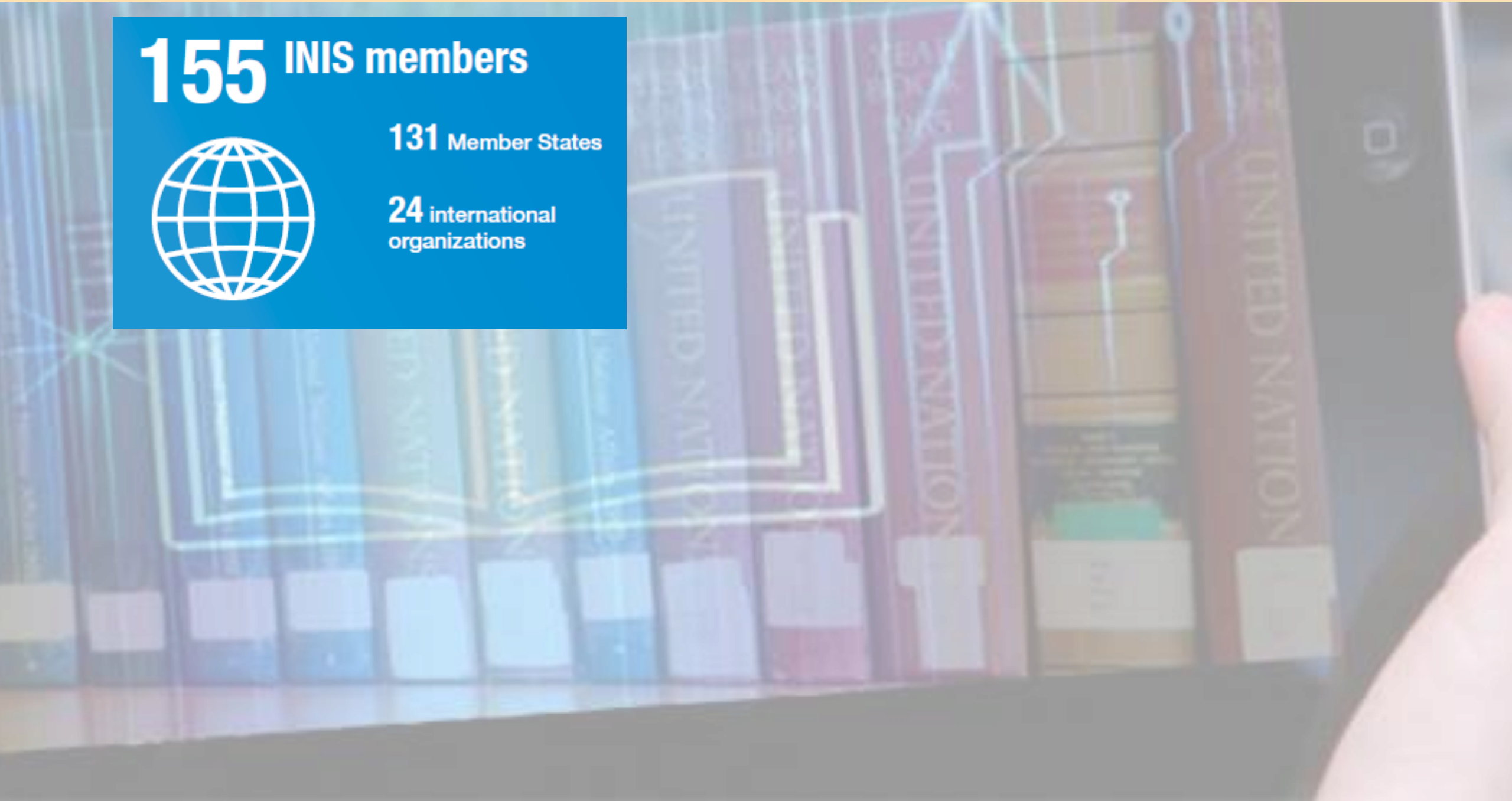




OUTLINE

- Introduction
- International Nuclear Information System
- INIS workflow
- Challenges
- Two-pass indexing
- Results and conclusions
- Outcomes and future work

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

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

- **4.2 million** bibliographic records
  108 000 in 2018
- **570 000** full-text documents
  19 500 in 2018
- **3.2 million** web page views
- **1.6 million** unique searches
- **1 million** unique visitors

inis.iaea.org

inis.iaea.org/search

INIS thesaurus

A knowledge organization system

8 languages

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

31 309 descriptors

150 in 2018

INIS THESAURUS

INIS thesaurus

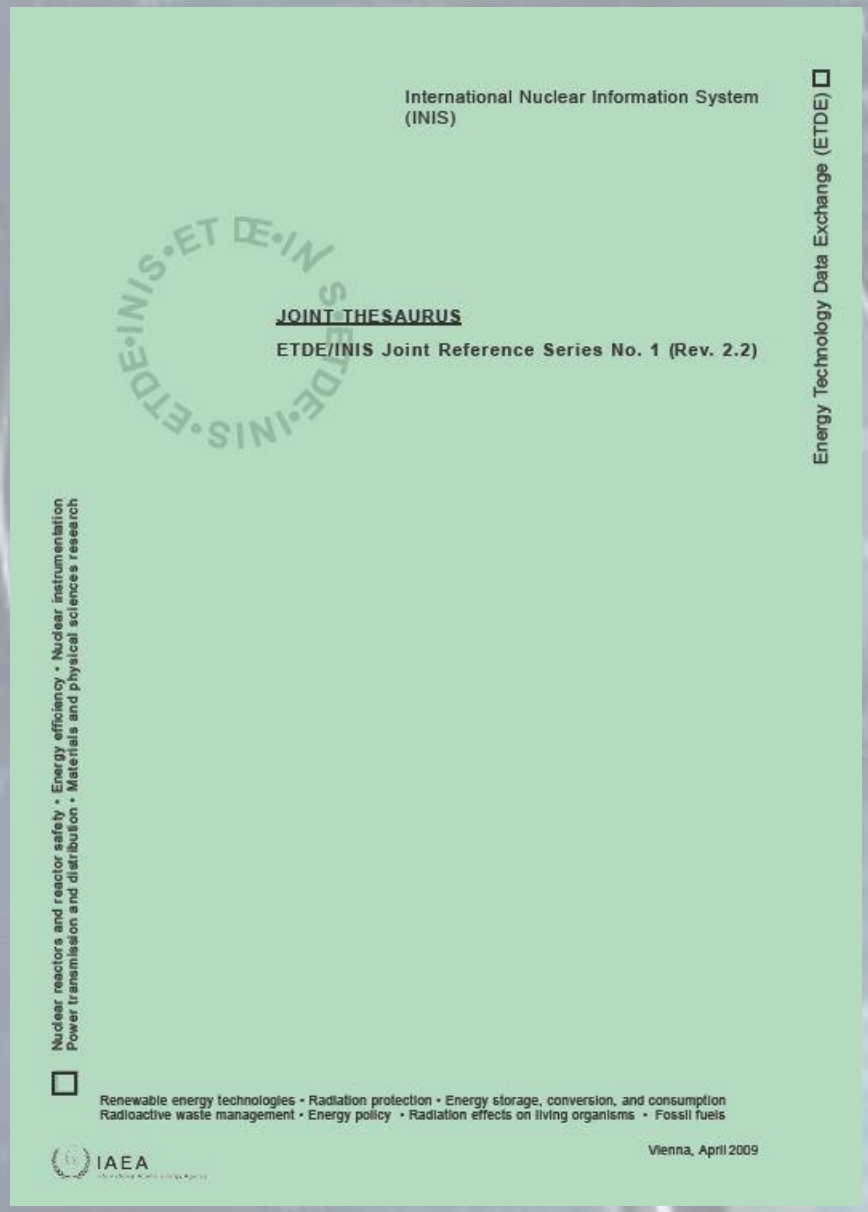
A knowledge organization system

8 languages

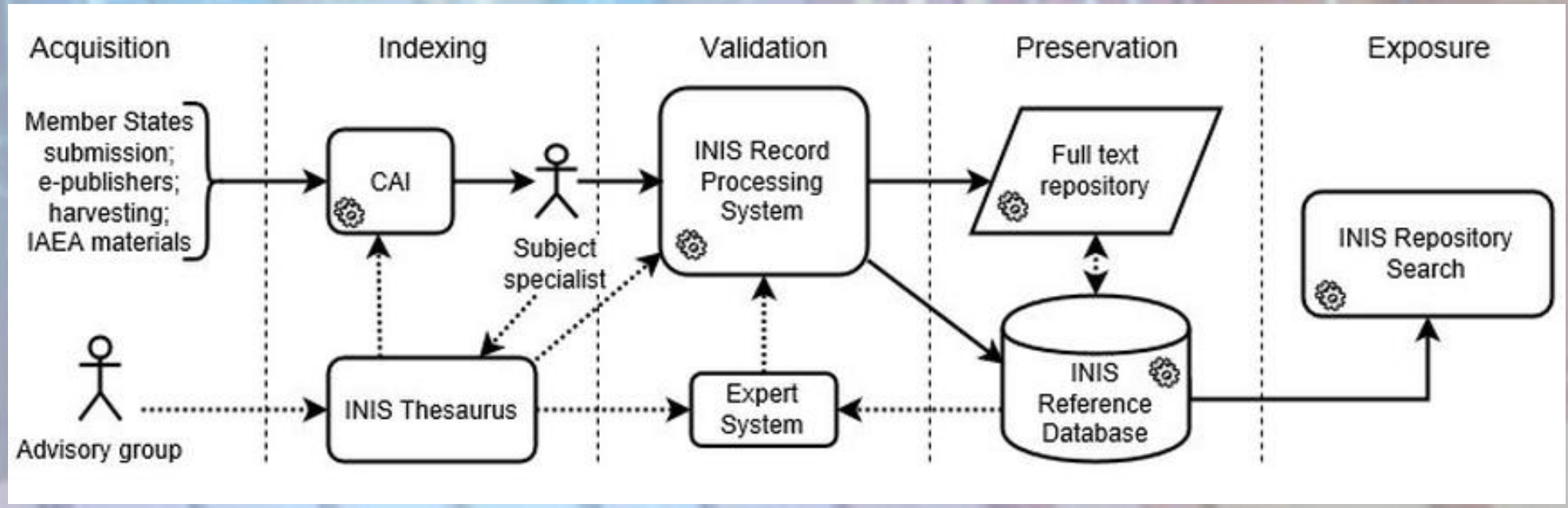
Arabic, Chinese, English, French,
German, Japanese, Russian, Spanish

31 309 descriptors

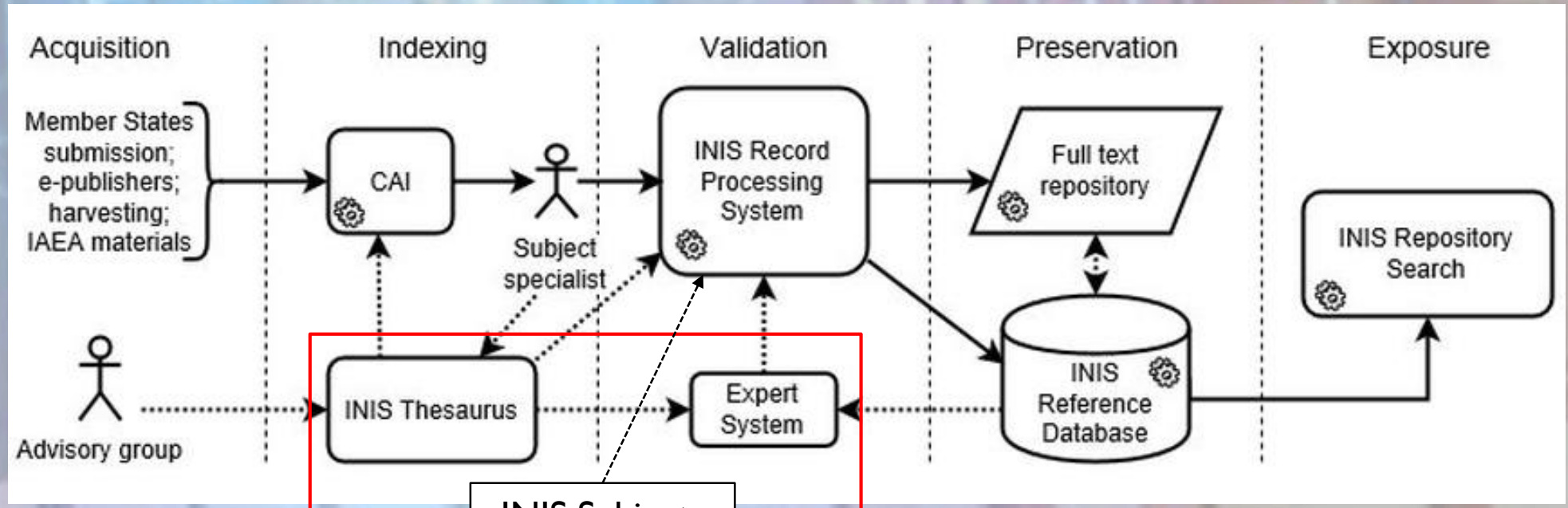
150 in 2018



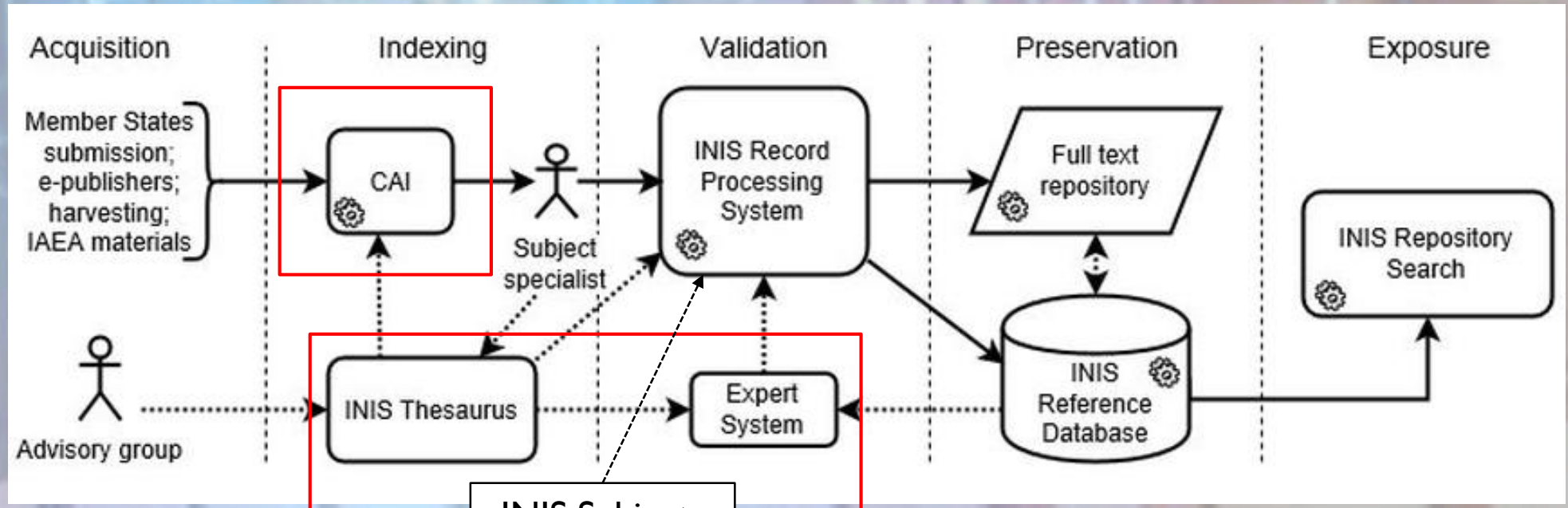
INIS WORKFLOW



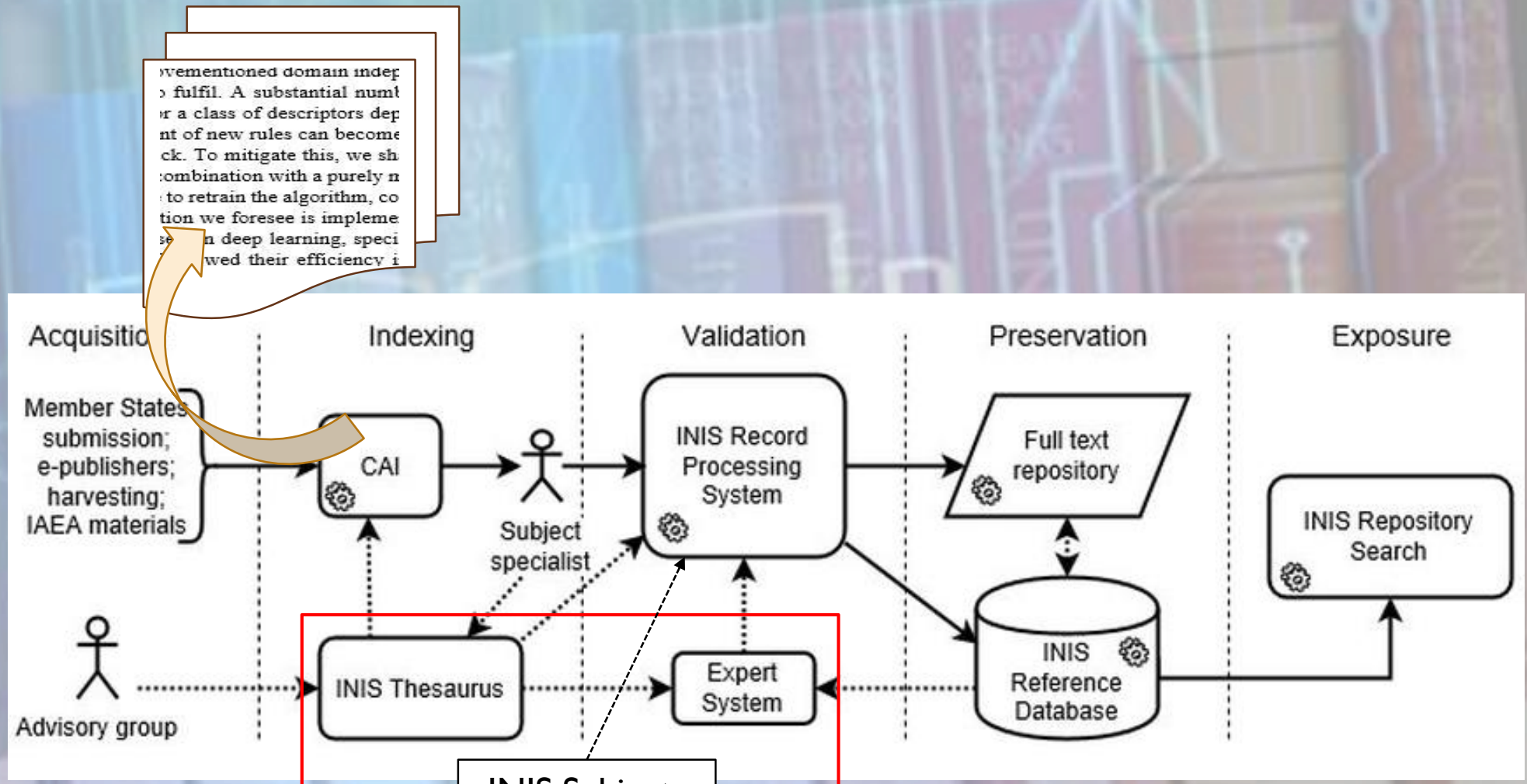
INIS WORKFLOW



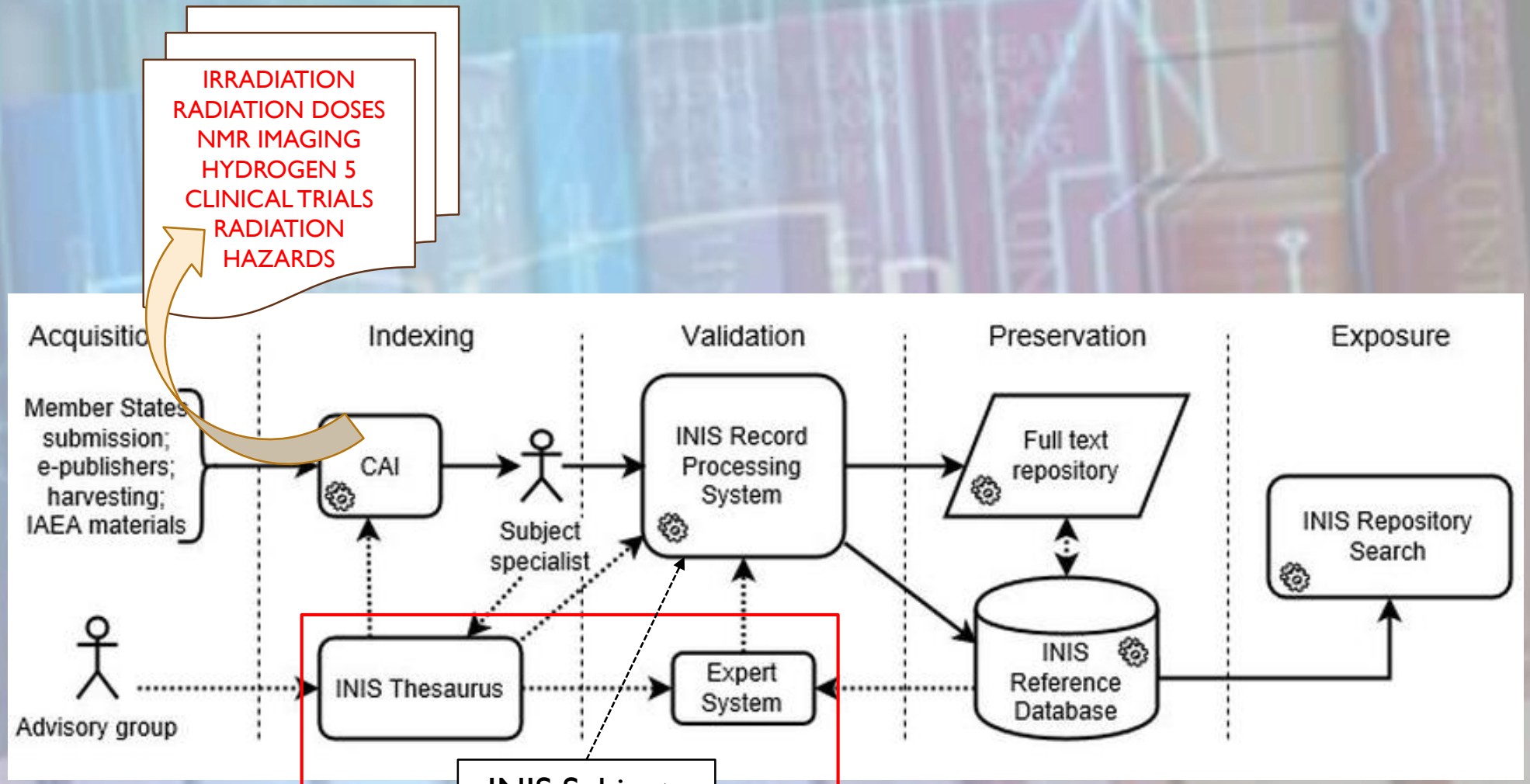
INIS WORKFLOW



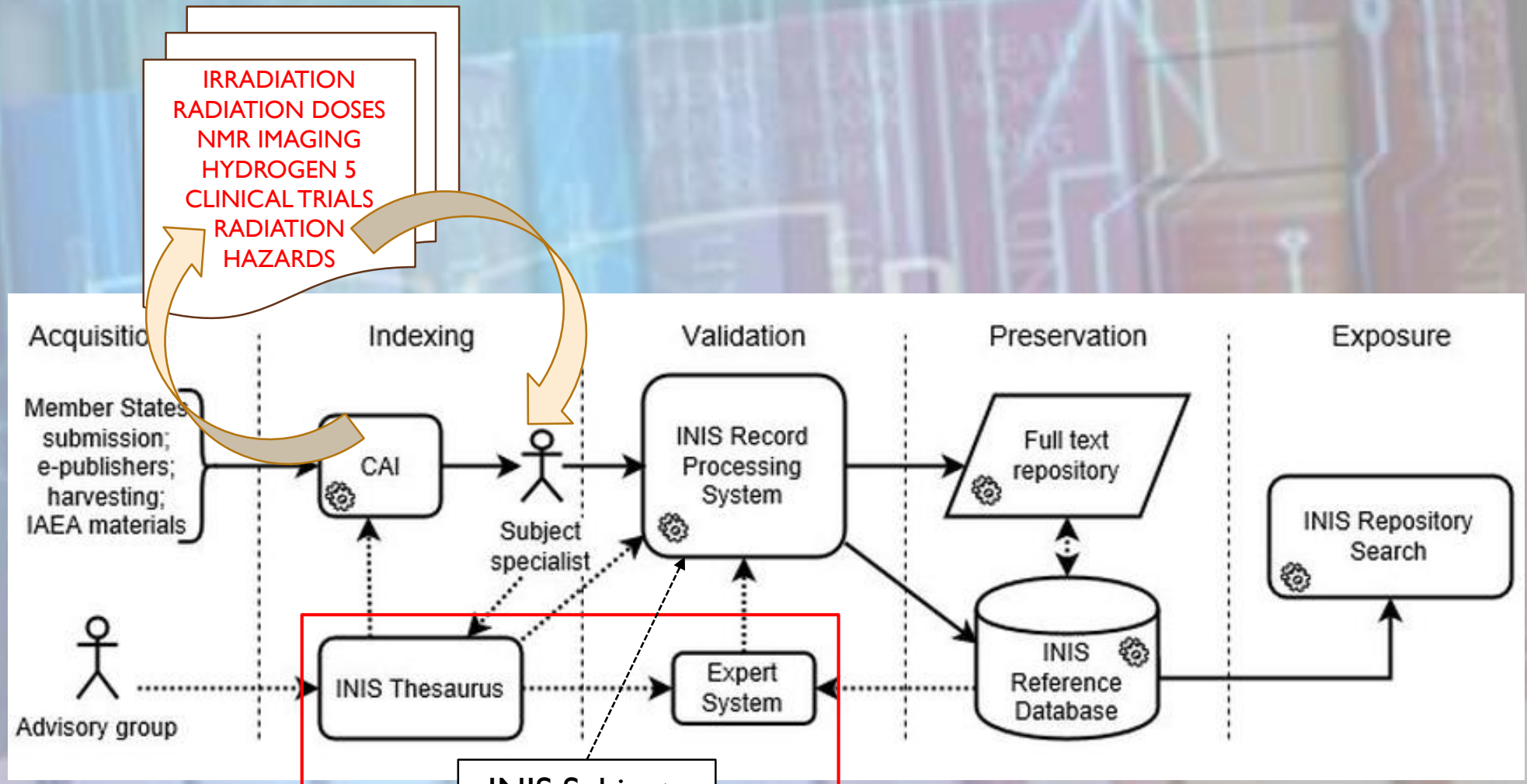
INIS WORKFLOW



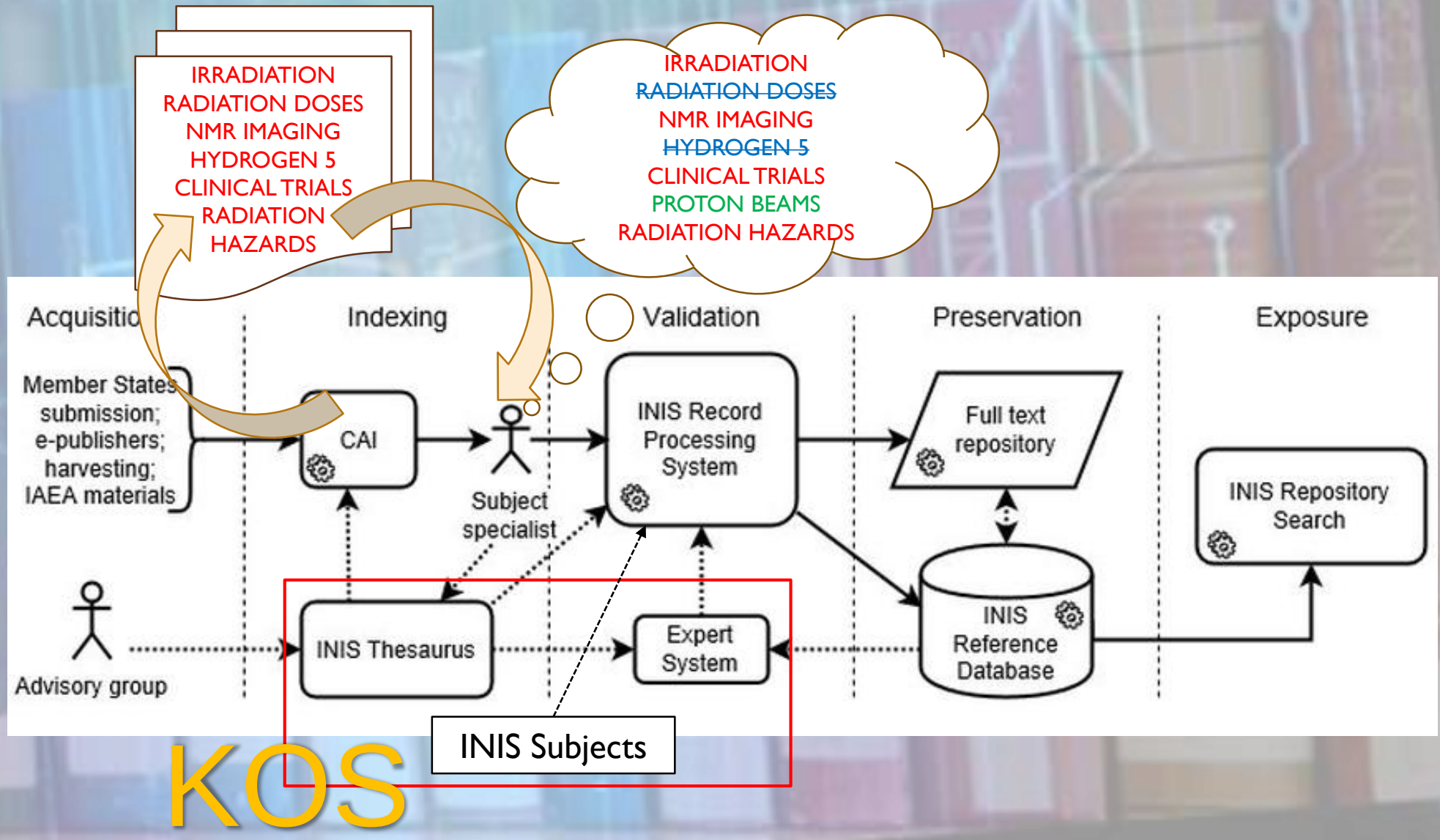
INIS WORKFLOW



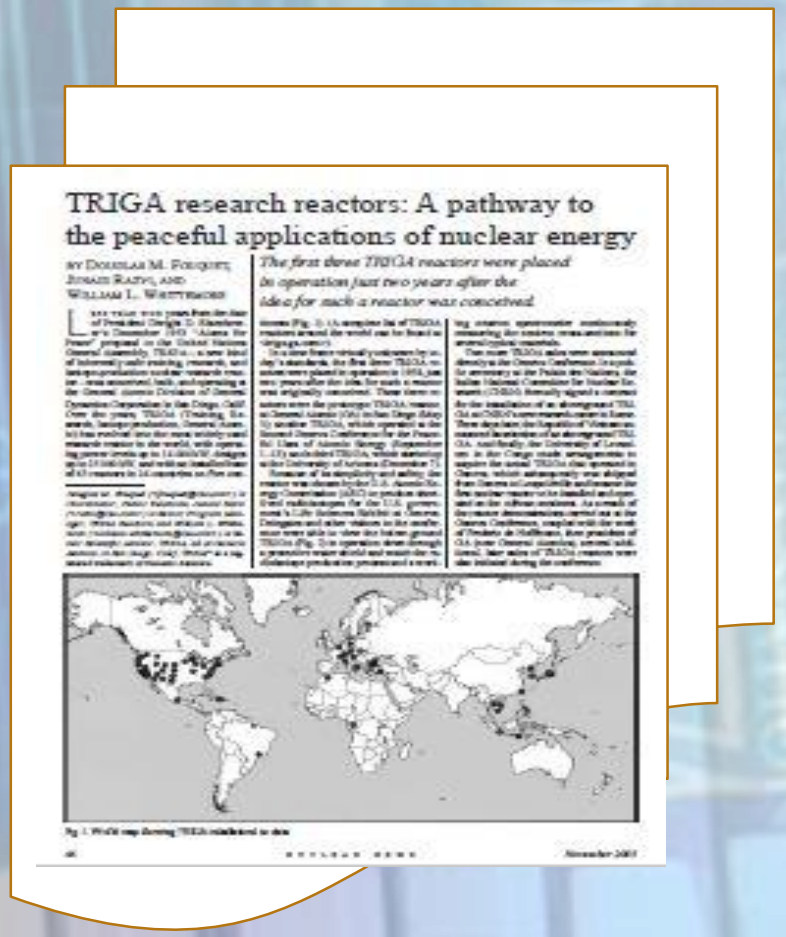
INIS WORKFLOW



INIS WORKFLOW



CHALLENGES



TRIGA research reactors: A pathway to the peaceful applications of nuclear energy

The first three TRIGA reactors were placed in operation just two years after the idea for such a reactor was conceived.

CHALLENGES

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

CHALLENGES

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

Resulting subject analysis often contains

CHALLENGES

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

Resulting subject analysis often contains

- Too broad descriptors

CHALLENGES

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

"…leads to a melting"

Resulting subject analysis often contains

- Too broad descriptors
- Misleading suggestions

CHALLENGES

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

Resulting subject analysis often contains

- Too broad descriptors
- Misleading suggestions
- Incorrect interpretations

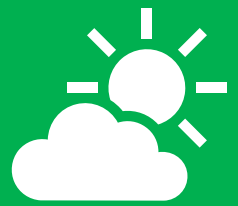
CHALLENGES

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

Resulting subject analysis often contains

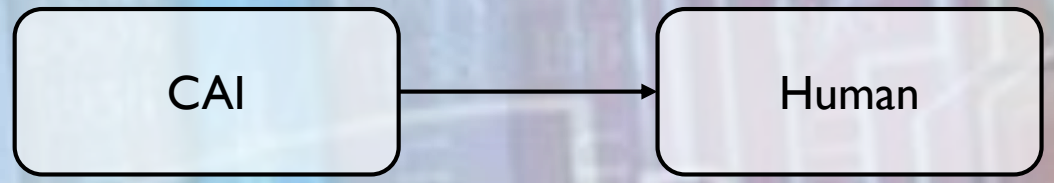
- Too broad descriptors
- Misleading suggestions
- Incorrect interpretations

50 years of INIS

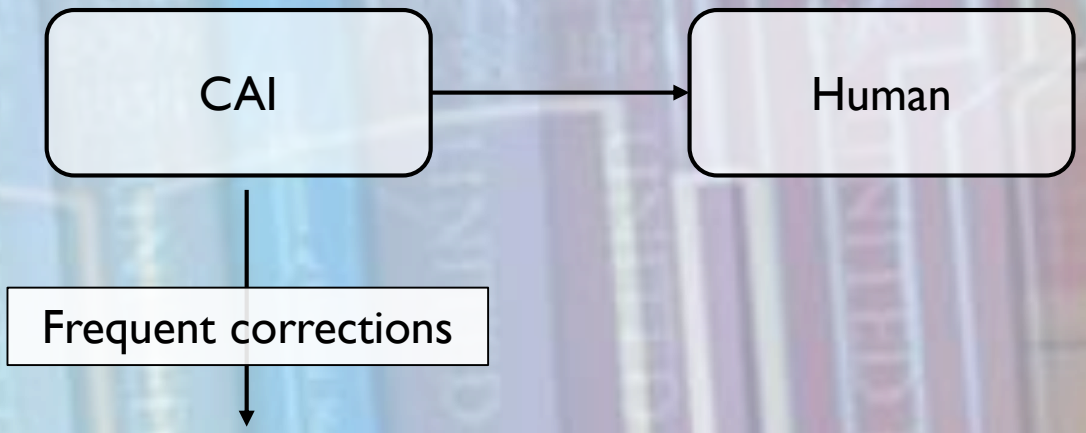


- 4 million indexed records
- Experienced specialists
- Unique KOS and workflow

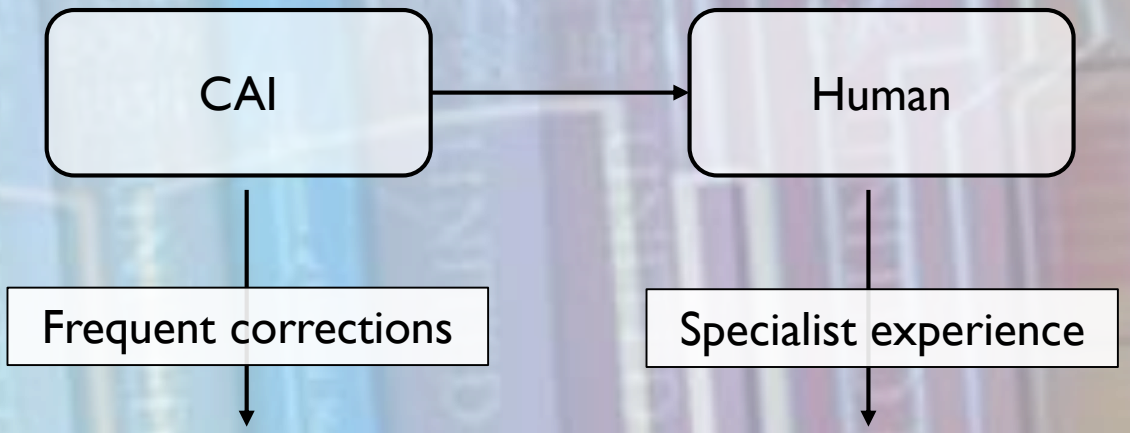
TWO-PASS INDEXING, TIER 1



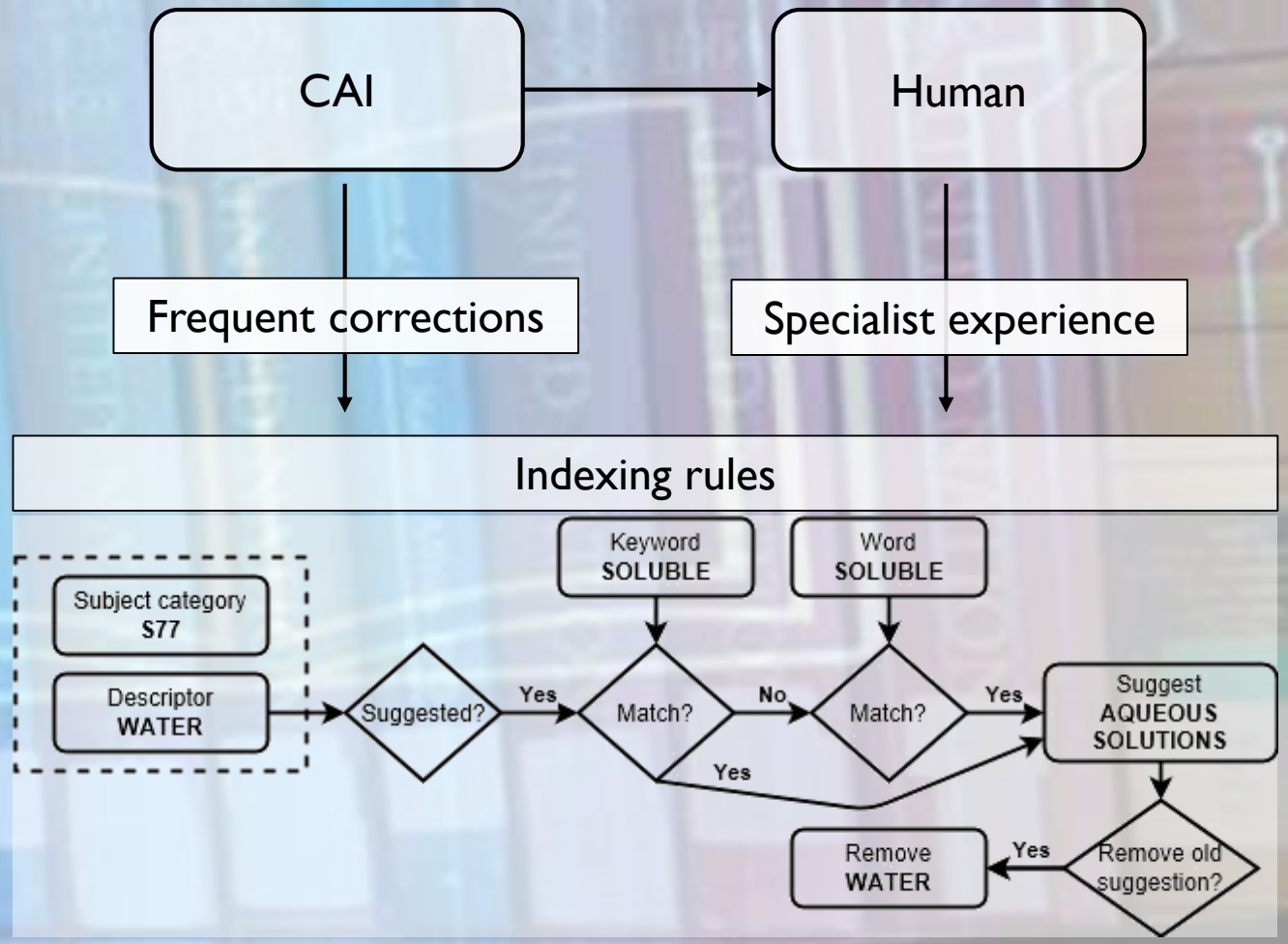
TWO-PASS INDEXING, TIER 1



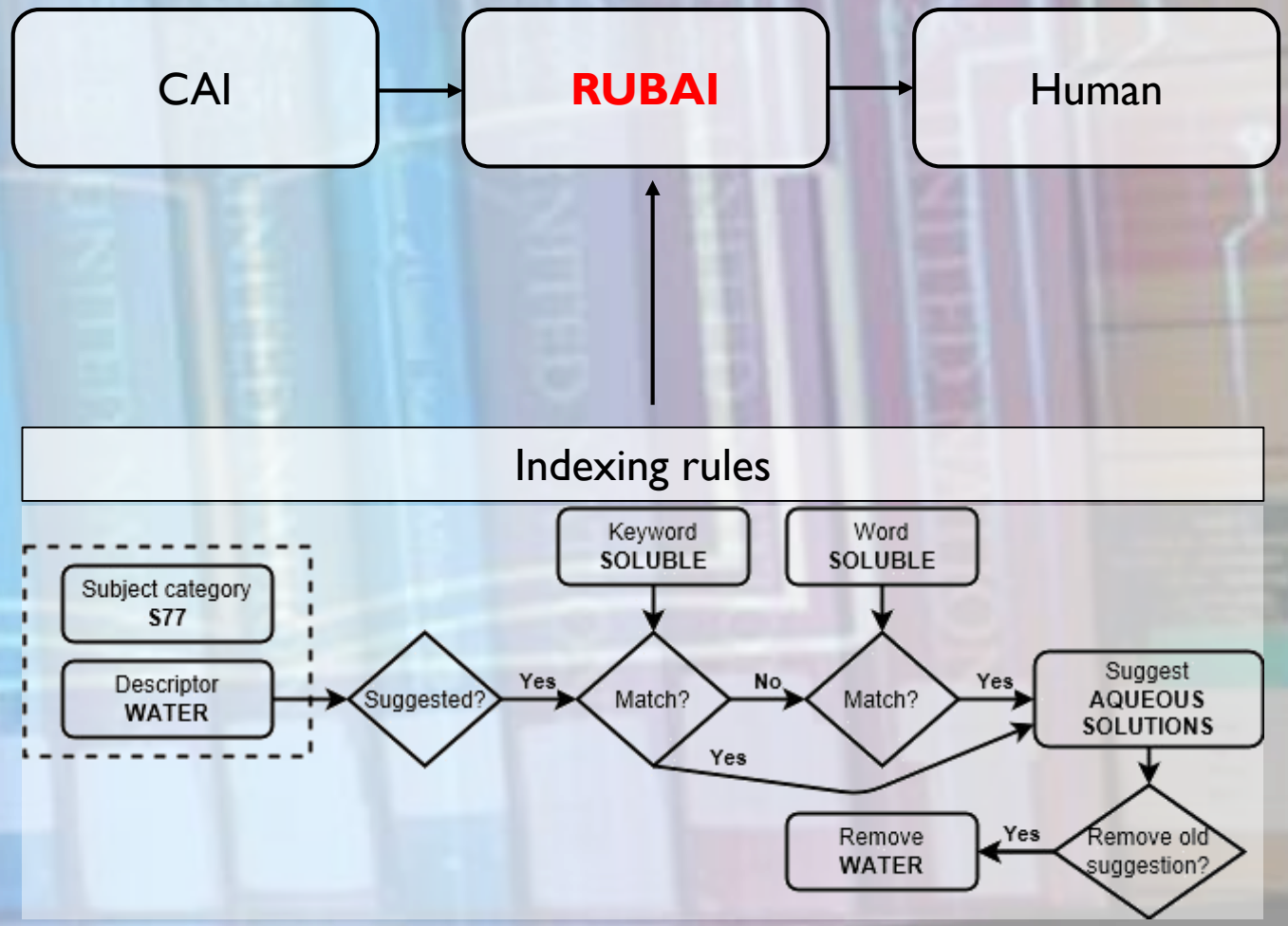
TWO-PASS INDEXING, TIER 1



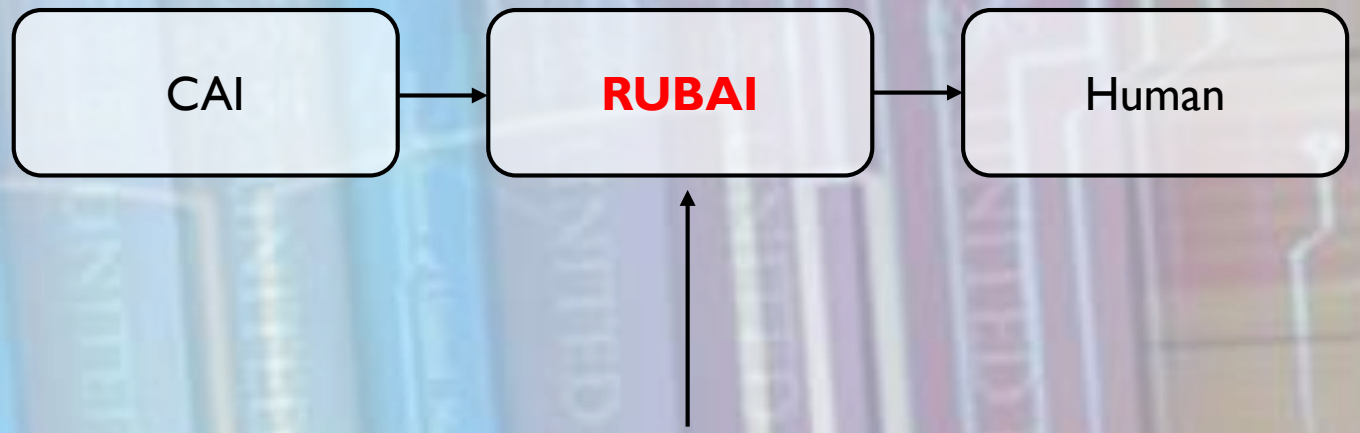
TWO-PASS INDEXING, TIER 1



TWO-PASS INDEXING, TIER I



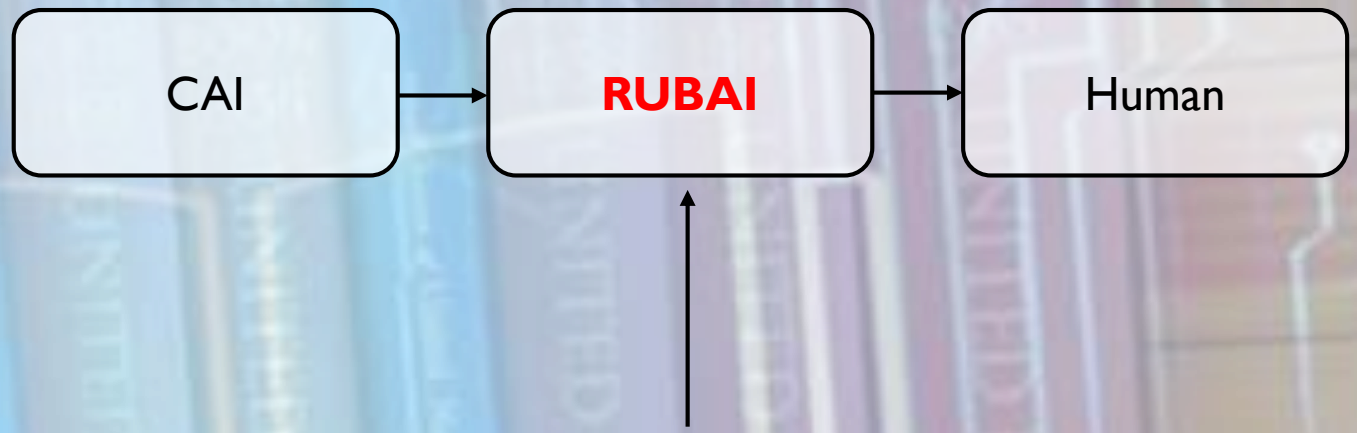
TWO-PASS INDEXING, TIER I



940 unique rules

- Add
- Replace
- Remove
- Keep

TWO-PASS INDEXING, TIER I



940 unique rules

- Add
- Replace
- Remove
- Keep

KOS integration

- Thesaurus
- INIS subjects
- Expert knowledge
- Quality control

RUBAI: RESULTS

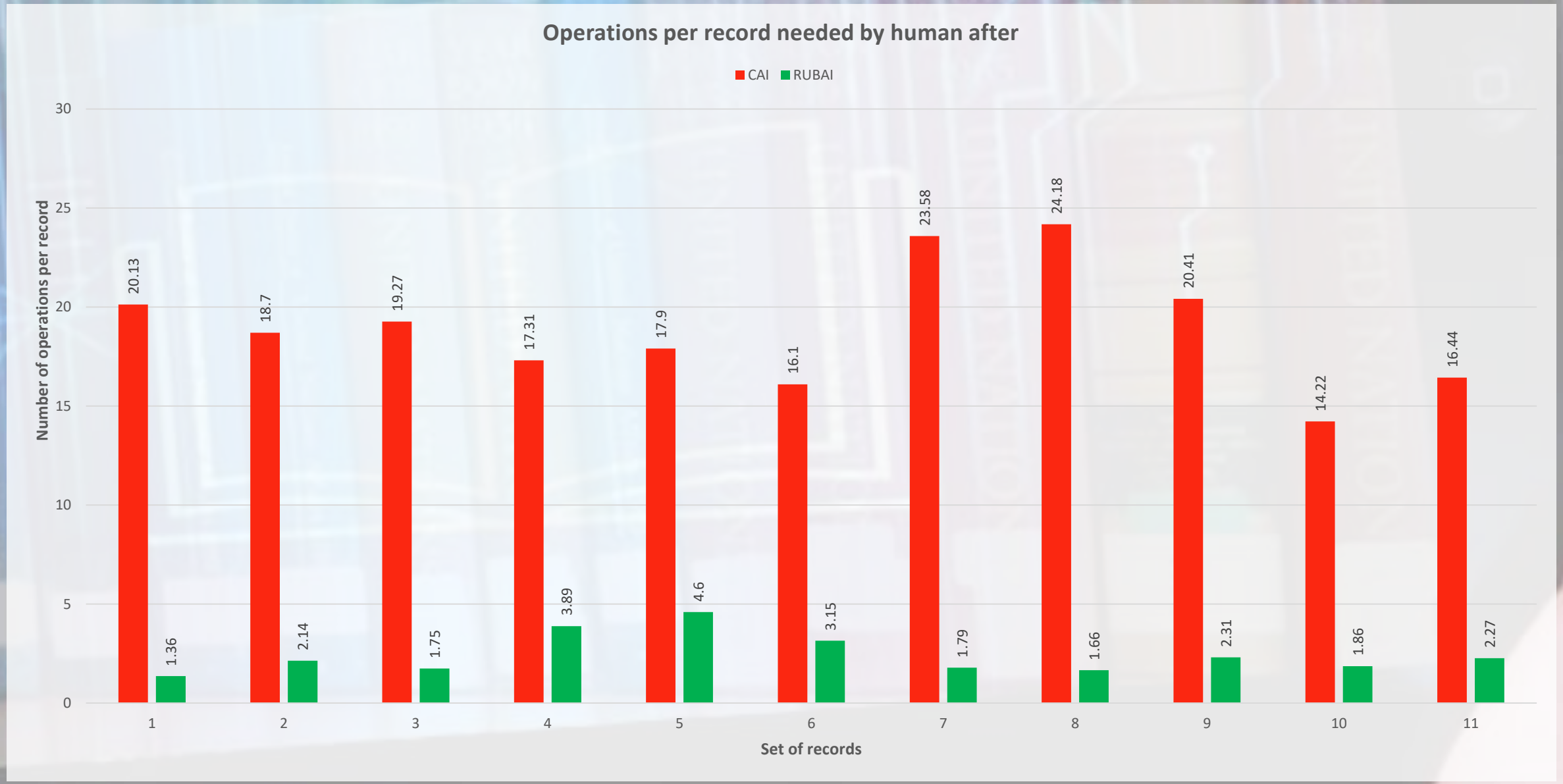
ID	Number of records	Descriptors							Records where more than one descriptor were				Operations per record performed by human after…	
		Total after…			added by human after…		removed by human after…		added after…		removed after…			
		CAI	RUBAI	Human	CAI	RUBAI	CAI	RUBAI	CAI	RUBAI	CAI	RUBAI	CAI	RUBAI
1	94	2636	1154	1032	144	3	1748	125	42	0	94	36	***20.13***	***1.36***
2	94	2570	1177	984	86	4	1672	197	24	1	94	67	***18.7***	***2.14***
3	97	2473	875	845	121	70	1749	100	37	18	97	32	***19.27***	***1.75***
4	95	2361	900	900	92	185	1553	185	19	46	95	49	***17.31***	***3.89***
5	96	2398	1154	916	118	102	1600	340	27	28	96	76	***17.9***	***4.6***
6	91	2166	1050	929	114	83	1351	204	26	24	91	52	***16.1***	***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

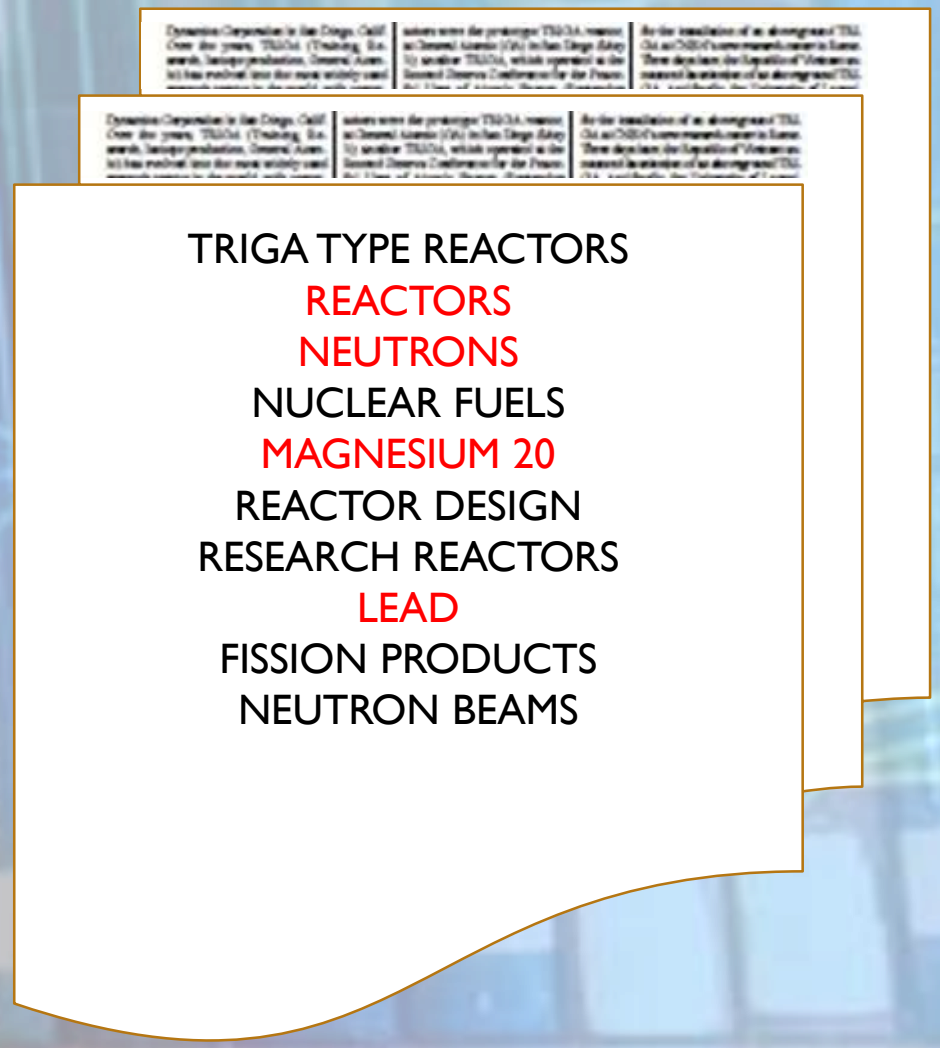
ID	Number of records	Descriptors							Records where more than one descriptor were				Operations per record performed by human after…	
		Total after…			added by human		removed by human		added after…		removed after…			
														UBAI
1														*1.36*
2														*2.14*
3														*1.75*
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***

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

RUBAI: RESULTS

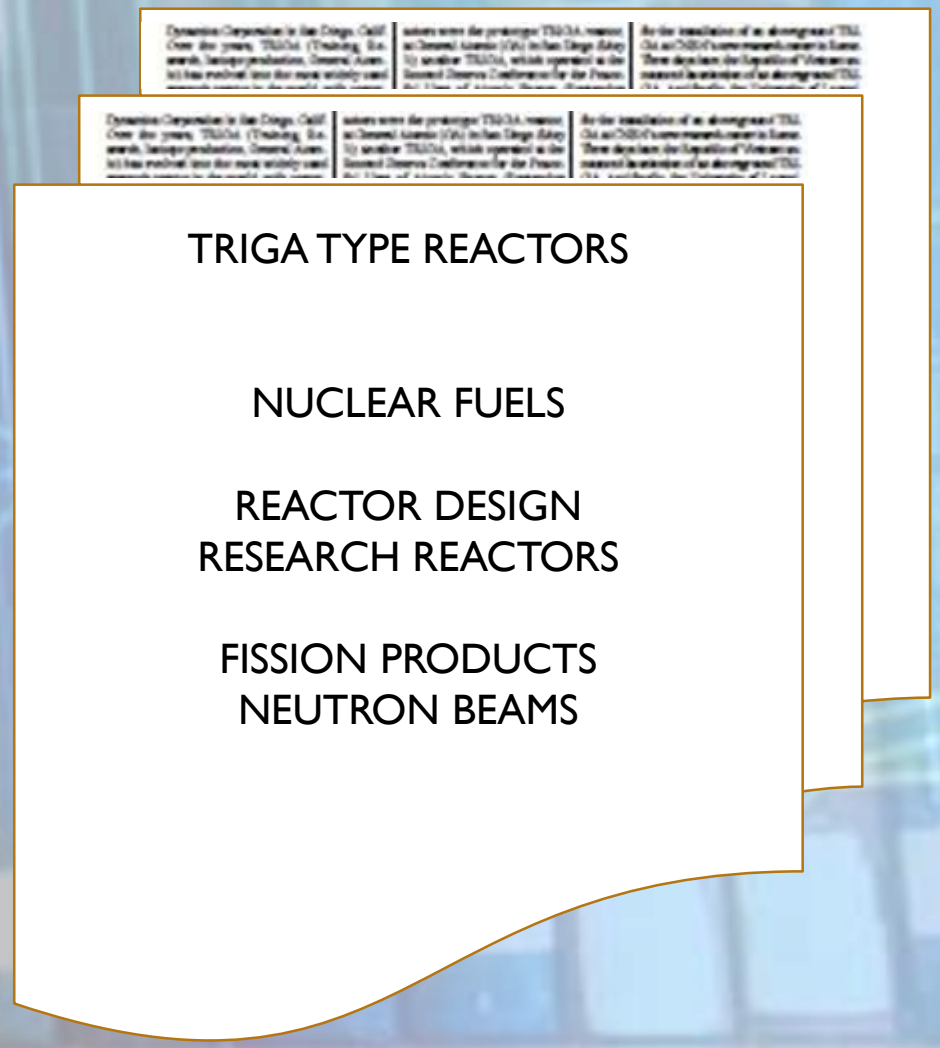


ARE WE DOING WELL?



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

ARE WE DOING WELL?



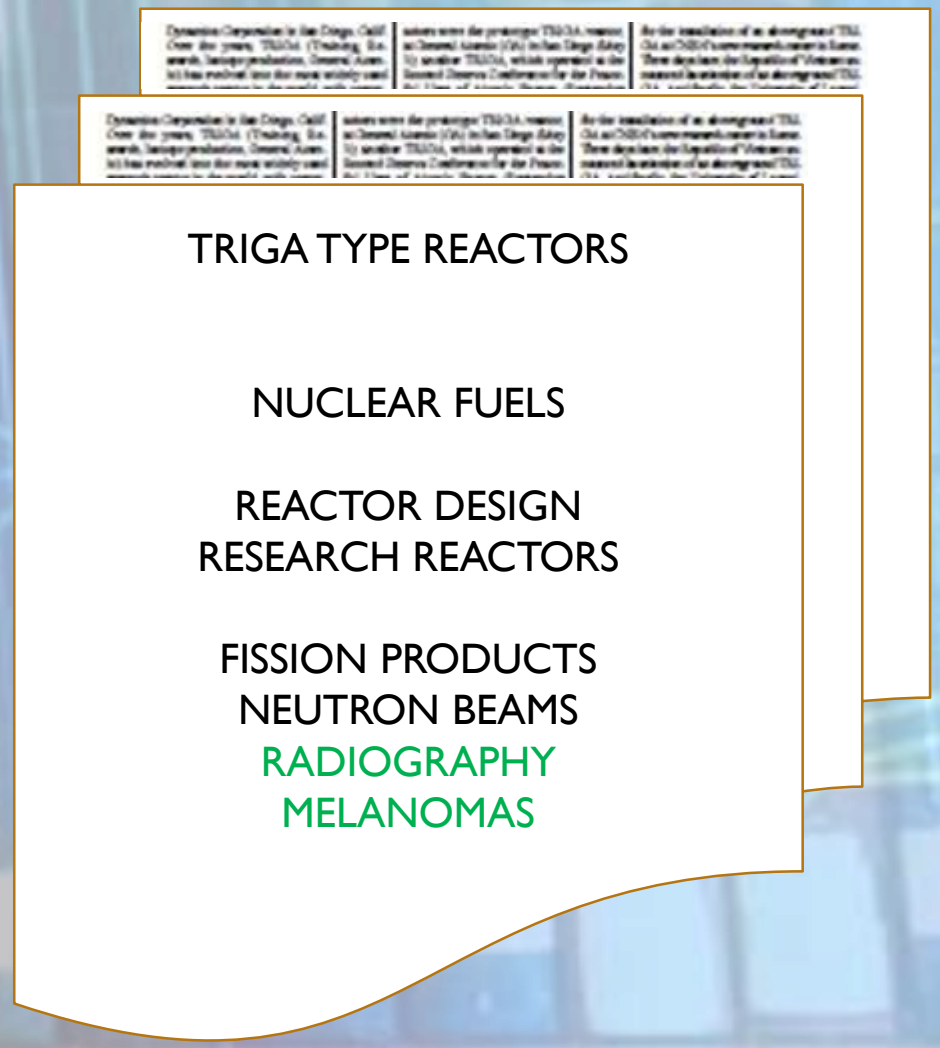
TRIGA TYPE REACTORS

NUCLEAR FUELS

REACTOR DESIGN
RESEARCH REACTORS

FISSION PRODUCTS
NEUTRON BEAMS

ARE WE DOING WELL?



TRIGA TYPE REACTORS

NUCLEAR FUELS

REACTOR DESIGN
RESEARCH REACTORS

FISSION PRODUCTS
NEUTRON BEAMS
RADIOGRAPHY
MELANOMAS

ARE WE DOING WELL?

TRIGA TYPE REACTORS

NUCLEAR FUELS

REACTOR DESIGN
RESEARCH REACTORS

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

Descriptor	RUBAI Operation	Human Action	Human Decision
TRIGA TYPE REACTORS	Keep	Keep	Approve
REACTORS	Remove		Approve
RESEARCH REACTORS	Keep	Keep	Approve
LEAD	Remove		Approve
RADIOGRAPHY	Add	Remove	Reject
MELANOMAS	Add	Remove	Reject
NEUTRON CAPTURE THERAPY		Add	
ISOTOPE PRODUCTION		Add	
INDUSTRIAL RADIOGRAPHY		Add	

FROM A SINGLE DECISION TO A DATASET

Descriptor	RUBAI Operation	Human Action	Human Decision
TRIGA TYPE REACTORS	Keep	Keep	Approve
REACTORS	Remove		Approve
RESEARCH REACTORS	Keep	Keep	Approve
LEAD	Remove		Approve
RADIOGRAPHY	Add	Remove	Reject
MELANOMAS	Add	Remove	Reject
NEUTRON CAPTURE THERAPY	Not Add	Add	
ISOTOPE PRODUCTION	Not Add	Add	
INDUSTRIAL RADIOGRAPHY	Not Add	Add	

FROM A SINGLE DECISION TO A DATASET

Descriptor	RUBAI Operation	Human Action	Human Decision
TRIGA TYPE REACTORS	Keep	Keep	Approve
REACTORS	Remove		Approve
RESEARCH REACTORS	Keep	Keep	Approve
LEAD	Remove		Approve
RADIOGRAPHY	Add	Remove	Reject
MELANOMAS	Add	Remove	Reject
NEUTRON CAPTURE THERAPY	Not Add	Add	Reject
ISOTOPE PRODUCTION	Not Add	Add	Reject
INDUSTRIAL RADIOGRAPHY	Not Add	Add	Reject

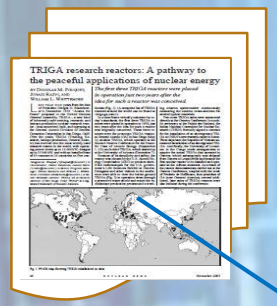
FROM A SINGLE DECISION TO A DATASET

Descriptor	RUBAI Operation	Human Action	Human Decision
TRIGA TYPE REACTORS	Keep	Keep	Approve
REACTORS	Remove		Approve
RESEARCH REACTORS	Keep	Keep	Approve
LEAD	Remove		Approve
RADIOGRAPHY	Add	Remove	Reject
MELANOMAS	Add	Remove	Reject

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

FROM A SINGLE DECISION TO A DATASET



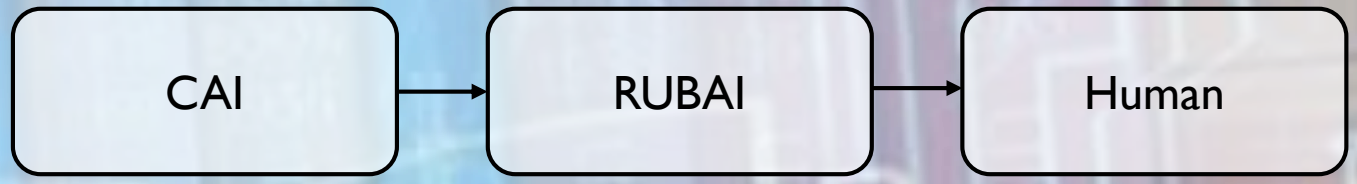
+

Descriptor	RUBAI Operation	Human Action	Human Decision
TRIGA TYPE REACTORS	Keep	Keep	Approve
REACTORS	Remove		Approve
RESEARCH REACTORS	Keep	Keep	Approve
LEAD	Remove		Approve
RADIOGRAPHY	Add	Remove	Reject
MELANOMAS	Add	Remove	Reject

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



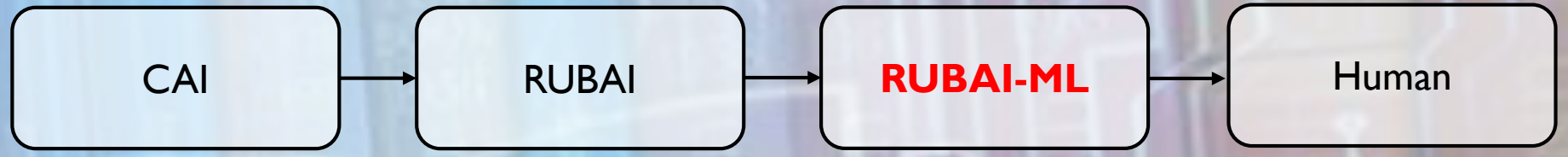
840 unique rules

- Add
- Replace
- Remove
- Keep

KOS integration

- Thesaurus
- INIS subjects
- Expert knowledge
- Quality control

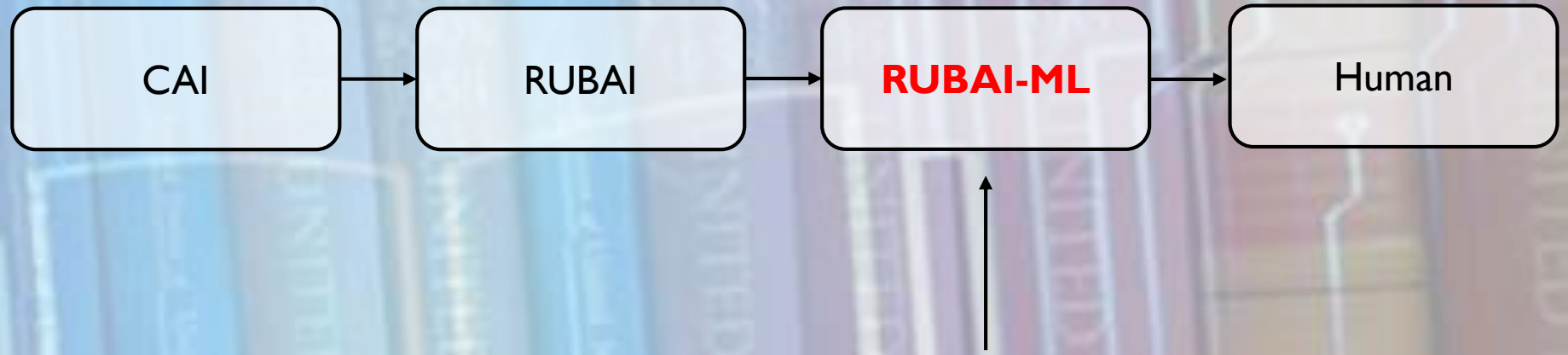
TWO-PASS INDEXING, TIER 2



840 KOS integration

- Thesaurus
- INIS subjects
- Expert knowledge
- Quality control

TWO-PASS INDEXING, TIER 2



840

KOS inte
The
INIS
Exp
Qua

DATA analysis
Machine Learning
KOS
Modeling human cognition

DATA set
17,500 documents
84% accuracy
130,000 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

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	Descriptors										Operations per record performed by human after…	
		Total after…				added by human after…			removed by human after…				
		CAI	RUBAI	RUBAI-ML	Human	CAI	RUBAI	RUBAI-ML	CAI	RUBAI	RUBAI-ML	CAI	RUBAI
1	120	2519	1032	1040	1094	205	262	257	1626	200	202	***3.85***	***3.83***
2	111	2104	859	862	936	175	232	231	1343	155	157	***3.49***	***3.49***
3	109	2217	916	915	960	160	216	220	1417	172	175	***3.56***	***3.62***
4	22	474	169	172	198	44	53	52	320	24	26	***3.5***	***3.54***

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…					Operations per record performed by human after…	
		CAI	RUBAI	RUBAI-M		RUBAI-ML	CAI	RUBAI
1	120	2519	1032	10	00	202	*3.85*	*3.83*
2	111	2104	859	8	55	157	*3.49*	*3.49*
3	109	2217	916	9	72	175	*3.56*	*3.62*
4	22	474	169	1	24	26	*3.5*	*3.54*

Operations per record performed by human after…

RUBAI	RUBAI-ML
3.60	3.62

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 descriptors after…				per record uman after…
		CAI	RUBAI	RUBAI-ML	HUMAN	RUBAI
1	120	2519	1032	1040	1094	*3.83*
2	111	2104	859	862	936	*3.49*
3	109	2217	916	915	960	*3.62*
4	22	474	169	172	198	*3.54*
		1829	**744**	**747**	**797**	

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 descriptors after…				per record human after…
		CAI	RUBAI	RUBAI-ML	HUMAN	RUBAI
1	120	2519	1032	1040	1094	*3.83*
2	111	2104	859	862	936	*3.49*
3	109	2217	916	915	960	*3.62*
4	22	474	169	172	198	*3.54*
		1829	**744**	**747**	**797**	

… CLOSER TO THE HUMAN CHOICES

Identifies missing descriptors in case of:

- Specific semantic relations
- Narrower descriptors
- Broader document scope

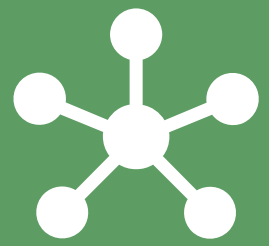
CONCLUSIONS

840 unique rules

- Add
- Replace
- Remove
- Keep

CONCLUSIONS

840

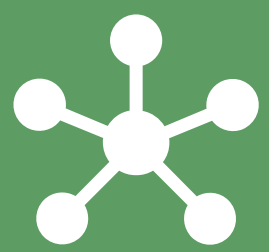


KOS integration

- Thesaurus
- INIS subjects
- Expert knowledge
- Quality control

CONCLUSIONS

840



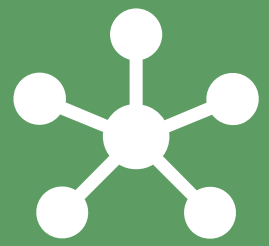
KOS

DATA analysis

- Machine Learning
- KOS
- Modeling human cognition

CONCLUSIONS

CONCLUSIONS

840 

KOS

DATA

DATA set

- 17,500 documents
- 84% accuracy
- 130,000 unique decisions

Work experience
Expert knowledge

+

KOS integration

+

Data analysis
Decision simulation

Main outcomes:

- Model human decision-making strategies
- Gradual optimization of the indexing quality
- Automating the indexing process

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

THANK YOU !

International Nuclear Information System (INIS)
inis.iaea.org

INIS Repository
inis.iaea.org/search

INIS: The First Forty Years
www.iaea.org/sites/default/files/inis-40-anniversary.pdf