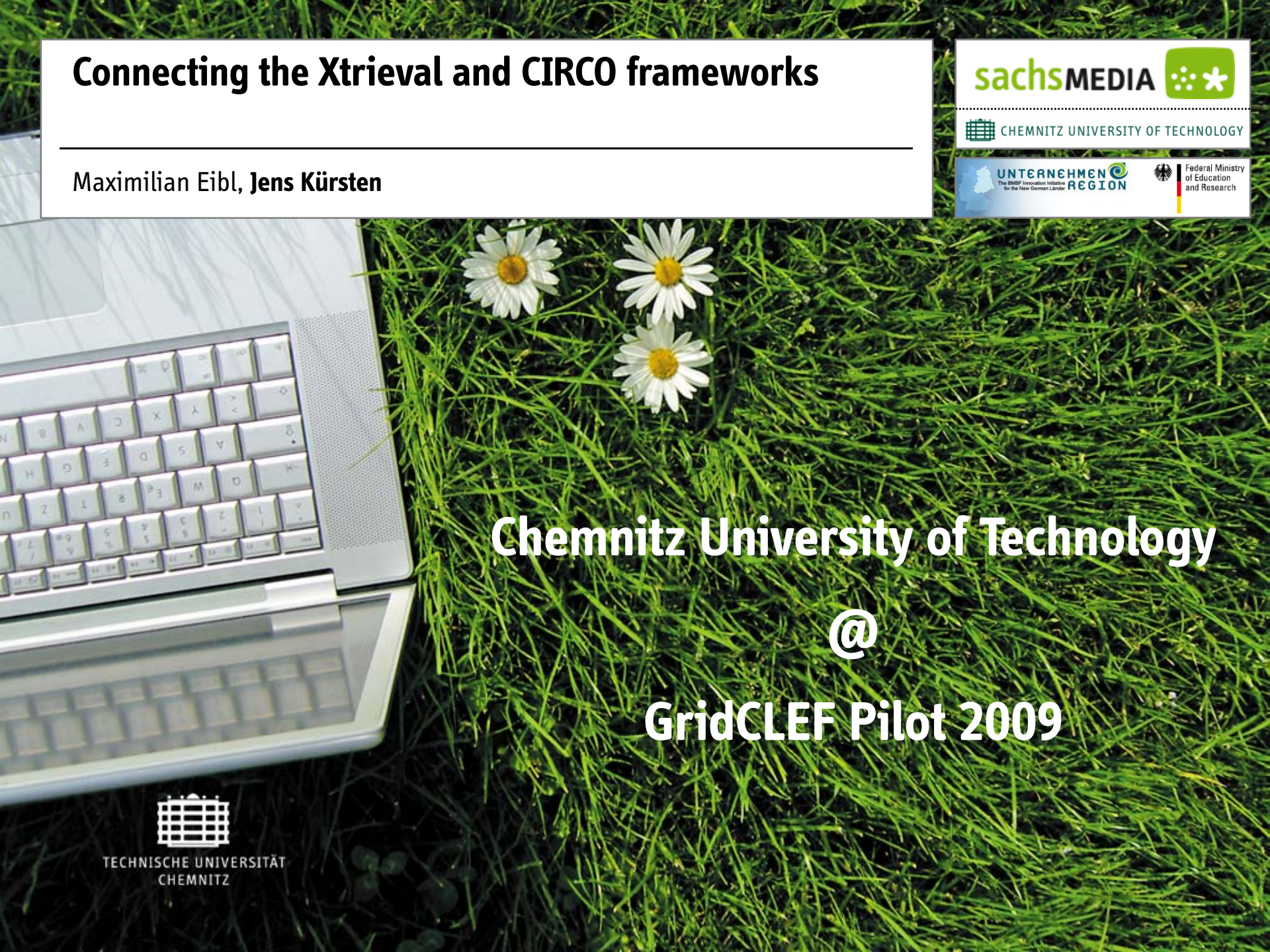
Connecting the Xtrieval and CIRCO frameworks

Maximilian Eibl, **Jens Kürsten**

sachsMEDIA
CHEMNITZ UNIVERSITY OF TECHNOLOGY
UNTERNEHMEN REGION
The BMBF Innovation Initiative for the New German Länder
Federal Ministry of Education and Research



Chemnitz University of Technology

@

GridCLEF Pilot 2009

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Outline

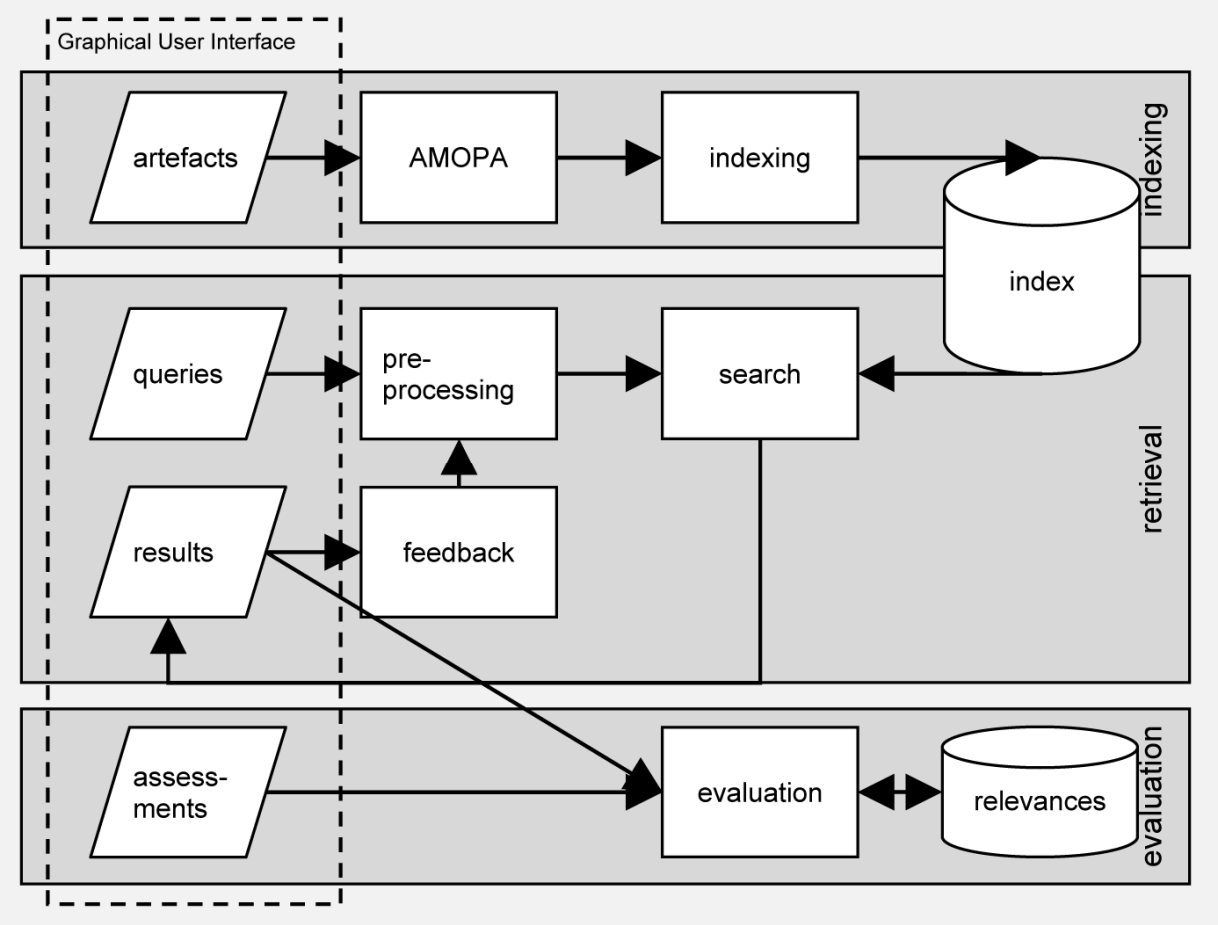
- Motivation
- Integrating CIRCO in Xtrieval
- Experimental results and analysis
- Lessons learned
- Conclusion and future work

Xtrieval within sachsMedia research project

sachsMEDIA

CHEMNITZ UNIVERSITY OF TECHNOLOGY

- sachsMedia: towards a TV archive for collaboration



Integrating CIRCO in Xtrieval

- Xtrieval: JAVA wrapper to use, compare and combine well known IR core toolkits (Lucene, Lemur, Terrier)
- easy CIRCO integration:
  - written in JAVA
  - was developed based on experiences with Lucene API
  - 4 additional lines of code for indexing procedure

Experimental results

sachsMEDIA

CHEMNITZ UNIVERSITY OF TECHNOLOGY

ID	Lang	Core	IR Model	Stemmer	# QE docs/terms	MAP
CUT_de_1	DE	Lucene	VSM	Snowball	10/50	0,4196
CUT_de_2	DE	Terrier	BM25	Snowball	10/50	0,4355
CUT_de_3	DE	Lucene	VSM	N-Gram	10/250	0,4267
CUT_de_4	DE	Terrier	BM25	N-Gram	10/250	0,4678
CUT_de_5	DE	both	both	both	10/50 & 250	0,4864
CUT_en_1	EN	Lucene	VSM	Snowball	10/20	0,5067
CUT_en_2	EN	Terrier	BM25	Snowball	10/20	0,4926
CUT_en_3	EN	Lucene	VSM	Krovetz	10/20	0,4937
CUT_en_4	EN	Terrier	BM25	Krovetz	10/20	0,4859
CUT_en_5	EN	both	both	both	10/20	0,5446
CUT_fr_3	FR	Lucene	VSM	Snowball	10/20	0,0025
CUT_fr_3*	FR	Lucene	VSM	Snowball	10/20	0,4483
CUT_fr_1	FR	Terrier	BM25	Snowball	10/20	0,4538
CUT_fr_5	FR	Lucene	VSM	Savoy	10/20	0,4434
CUT_fr_2	FR	Terrier	BM25	Savoy	10/20	0,4795
CUT_fr_4	FR	both	both	both	10/20	0,4942

UNTERNEHMEN REGION
The BMBF Innovation Initiative for the New German Länder
Federal Ministry of Education and Research

Result analysis – IR models

sachsMEDIA

CHEMNITZ UNIVERSITY OF TECHNOLOGY

ID	Lang	Core	IR Model	Stemmer	# QE docs/tokens	MAP
CUT_de_1	DE	Lucene	VSM	Snowball	10/50	0,4196
CUT_de_2	DE	Terrier	BM25	Snowball	10/50	0,4355
CUT_de_3	DE	Lucene	VSM	N-Gram	10/250	0,4267
CUT_de_4	DE	Terrier	BM25	N-Gram	10/250	0,4678
CUT_de_5	DE	both	both	both	10/50 & 250	0,4864
CUT_en_1	EN	Lucene	VSM	Snowball	10/20	0,5067
CUT_en_2	EN	Terrier	BM25	Snowball	10/20	0,4926
CUT_en_3	EN	Lucene	VSM	Krovetz	10/20	0,4937
CUT_en_4	EN	Terrier	BM25	Krovetz	10/20	0,4859
CUT_en_5	EN	both	both	both	10/20	0,5446
CUT_fr_3	FR	Lucene	VSM	Snowball	10/20	0,0025
CUT_fr_3*	FR	Lucene	VSM	Snowball	10/20	0,4483
CUT_fr_1	FR	Terrier	BM25	Snowball	10/20	0,4538
CUT_fr_5	FR	Lucene	VSM	Savoy	10/20	0,4434
CUT_fr_2	FR	Terrier	BM25	Savoy	10/20	0,4795
CUT_fr_4	FR	both	both	both	10/20	0,4942

Result analysis – Token processing

sachsMEDIA

CHEMNITZ UNIVERSITY OF TECHNOLOGY

ID	Lang	Core	IR Model	Stemmer	# QE docs/tokens	MAP
CUT_de_1	DE	Lucene	VSM	Snowball	10/50	0,4196
CUT_de_2	DE	Terrier	BM25	Snowball	10/50	0,4355
CUT_de_3	DE	Lucene	VSM	N-Gram	10/250	0,4267
CUT_de_4	DE	Terrier	BM25	N-Gram	10/250	0,4678
CUT_de_5	DE	both	both	both	10/50 & 250	0,4864
CUT_en_1	EN	Lucene	VSM	Snowball	10/20	0,5067
CUT_en_2	EN	Terrier	BM25	Snowball	10/20	0,4926
CUT_en_3	EN	Lucene	VSM	Krovetz	10/20	0,4937
CUT_en_4	EN	Terrier	BM25	Krovetz	10/20	0,4859
CUT_en_5	EN	both	both	both	10/20	0,5446
CUT_fr_3	FR	Lucene	VSM	Snowball	10/20	0,0025
CUT_fr_3*	FR	Lucene	VSM	Snowball	10/20	0,4483
CUT_fr_1	FR	Terrier	BM25	Snowball	10/20	0,4538
CUT_fr_5	FR	Lucene	VSM	Savoy	10/20	0,4434
CUT_fr_2	FR	Terrier	BM25	Savoy	10/20	0,4795
CUT_fr_4	FR	both	both	both	10/20	0,4942

Result analysis – Combination

sachsMEDIA

CHEMNITZ UNIVERSITY OF TECHNOLOGY

ID	Lang	Core	IR Model	Stemmer	# QE docs/tokens	MAP
CUT_de_1	DE	Lucene	VSM	Snowball	10/50	0,4196
CUT_de_2	DE	Terrier	BM25	Snowball	10/50	0,4355
CUT_de_3	DE	Lucene	VSM	N-Gram	10/250	0,4267
CUT_de_4	DE	Terrier	BM25	N-Gram	10/250	0,4678
CUT_de_5	DE	both	both	both	10/50 & 250	0,4864
CUT_en_1	EN	Lucene	VSM	Snowball	10/20	0,5067
CUT_en_2	EN	Terrier	BM25	Snowball	10/20	0,4926
CUT_en_3	EN	Lucene	VSM	Krovetz	10/20	0,4937
CUT_en_4	EN	Terrier	BM25	Krovetz	10/20	0,4859
CUT_en_5	EN	both	both	both	10/20	0,5446
CUT_fr_3	FR	Lucene	VSM	Snowball	10/20	0,0025
CUT_fr_3*	FR	Lucene	VSM	Snowball	10/20	0,4483
CUT_fr_1	FR	Terrier	BM25	Snowball	10/20	0,4538
CUT_fr_5	FR	Lucene	VSM	Savoy	10/20	0,4434
CUT_fr_2	FR	Terrier	BM25	Savoy	10/20	0,4795
CUT_fr_4	FR	both	both	both	10/20	0,4942

UNTERNEHMEN REGION
The BMBF Innovation Initiative for the New German Länder
Federal Ministry of Education and Research

Lessons learned

- very big XML files to process and exchange (maybe too large?)
- slows down processing ecpecially with compression
- protocol for element/attribute contents needed
- exchanging intermediate processing output needed?
- performance comparable to results from 2001/2002

BUT: only because of combination of different token processing and different IR models used!!!

Conclusion and future work

- Conclusion
  - CIRCO framework integrated
  - huge data processing output
  - alternative: exchanging code instead of data?
  - refining protocol
- Future work
  - test evaluation with Cheshire output !
  - identify system components to exchange

Q & A

- Thank you!
- Questions, answers and discussion