MTAS Workshop

Meertens Institute Amsterdam

Sept 25,26 2017

Goals for the Workshop

- Inform about the current state of MTAS
  - Applications of MTAS: Nederlab, FLAT
  - Experiences (of non-Meertens parties)
- MTAS current and future scope clarification
- Future developments
  - Collaborations
  - Flexible extensible architecture
- Embedding within CLARIN

Scope (1)

- Easy use & merging with new and existing corpus exploitation systems
- Operational environment:
  - Easy addition new data (SOLR)
  - Easy addition new formats
- Scalability:
  - 15 million documents and 9.5 billion words (NederLab)

Scope (2)

- Original focus on text corpora, linguistically enriched, provisions for chapter/paragraph structure
- Easily adapted for other text structures with a definite serial text token order
- But some challenges wrt.
  - transcribed media i.e. overlapping text segments
  - Hierarchical structures (not covered by CQL)
  - Subtokenizations: phonetics, morphology

Location: room 1.01 - *Spinhuis, commissariskamer*

Time	Title	Description	Person
12.00 - 13.00	*Lunch*		
12.45 - 13.00	**Welcome**	Introduction	Daan Broeder *Meertens Institute*
13.00 - 13.45	**Mtas Basics**	Background and theoretical basis; differences and interoperability with other search engines	Matthijs Brouwer *Meertens Institute*
13.45 - 14.15	**Experiences & requirements**	Nederlab project	Hennie Brugman *Meertens Institute*
14.15 - 14.45		Flat; Islandora/Fedora	Menzo Windhouwer *Meertens Institute*
14.45 - 15.15	*Coffee & Tea*		
15.15 - 15.45	**Experiences & requirements**	Experiences & Requirements ACDH	Matej Durco *OEAW Österreich*
15.45 - 16.15		Experiences & Requirements HZSK	Anne Freger Tommi Pirinen *Universität Hamburg*
16.15 - 16.45		Experiences & Requirements	Beto Boullosa *TU Darmstadt*
16.45 - …	**Discussion**		Daan Broeder *Meertens Institute*

Tuesday September 26

Location: room 2.18 - *Spinhuis*

Time	Title	Description	Name
09.00 - 09.30	**Practicalities**	Installation & configuration	Matthijs Brouwer *Meertens Institute*
09.30 - 10.00		User Interfaces	Hennie Brugman *Meertens Institute*
10.00 - 10.30		Programming with Mtas; Usage of the broker	Jan Pieter Kunst *Meertens Institute*
10.30 - 11.00	*Coffee & Tea*		
11.00 - 11.30	**Other**	CLARIN search engine strategy and Mtas	Daan Broeder *Meertens Institute*
11.30 - 12.30	**Discussion**	Mtas - future development plans & (development) collaboration	Daan Broeder *Meertens Institute*
12.30 - 13.00	*Lunch*		

The bigger picture

- Two strategies
- Aggregation
  - Aggregate all content at a central store
  - Normalize all the different formats
  - Provide a single unified exploitation environment
  - Pro: control, performance, effort, …
  - Con: IPR, scalability, …

The bigger picture

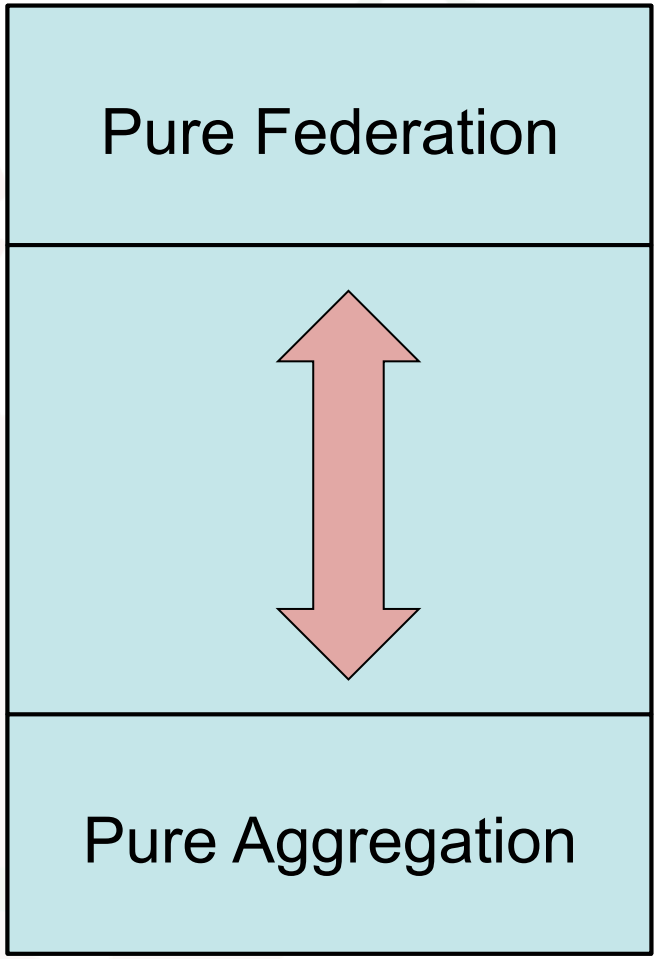
- Federation
  - Build a federation of search engines
  - Provide a portal normalizing queries & results
  - Pro: IPR, scalability, …
  - Con: query power & analysis, control, effort, …

CLARIN

- CLARIN supports FCS
- Currently limited functionality
- FCS is not a formal obligation, and difficult to get new federation members
- MTAS can be easily made FCS compliant

Should investigate alternatives

1. Federated MTAS
   - Allow more powerful queries
2. Index federation
3. Index aggregation



Collaboration options

- Coalition of the willing; MoU, looking for opportunities for extending & improving Mtas

- Central SCR; MI maintains core; contributions (configurations/parsers/…) from all
- Need code changes for more optimal collaboration?
  - Configuration & mapping files
  - Library of tokenizers/parser modules