Improving the Search Experience in a Social Network with Cross Media Contents

Daniele Cenni, Paolo Nesi

University of Florence
Department of Systems and Informatics
Distributed Systems and Internet Technology Laboratory

Paolo.nesi@unifi.it

cenni@dsi.unifi.it , http://www.disit.dinfo.unifi.it

ECLAP Social Network

- ECLAP is a **Digital Library** on Performing Arts connected with Europeana

- ECLAP is a **Best Practice and Social Network** (blogs, forums, comments, tagging, voting, …)

Goals/Requirements

- Develop an Indexing/Searching solution for ECLAP Social Network **allowing**:
  - Indexing **multilingual crossmedia** content metadata and data (e.g. documents)
  - Indexing portal blogs, forums, events, group pages, comments, etc.
  - **Efficient multilingual** search (keyword search and advanced search) supporting:
    - misspelled words (e.g. shespeare)
    - partial word search
  - Sorting and filtering search results
  - re-index the whole data without blocking the system
  - Log and monitor users activity
  - …
- Evaluate the Indexing/Searchig service

ECLAP ANY content kind

- **Informative Content**
  - Video, audio, images, documents
  - 3D, animations, Braille
  - Slide, Video-Slide, courses
  - eBook, ePub, Mpeg21, intelligent
- **Aggregated Content**:
  - Playlist, Collections
  - Annotations, Synchronization
- **Support and networking content**:
  - Blog, WebPage, Events, comments, forum, votes, messages, …

- Performance
- Master classes
- Scene Sketches
- Scenography
- Scenes
- Private lives of artists
- Scores
- Braille
- BackStage Stills
- Choreography
- Morals
- Poster
- Booklets
- Magazines Music
- Audio ballets

comments

rating

relationships

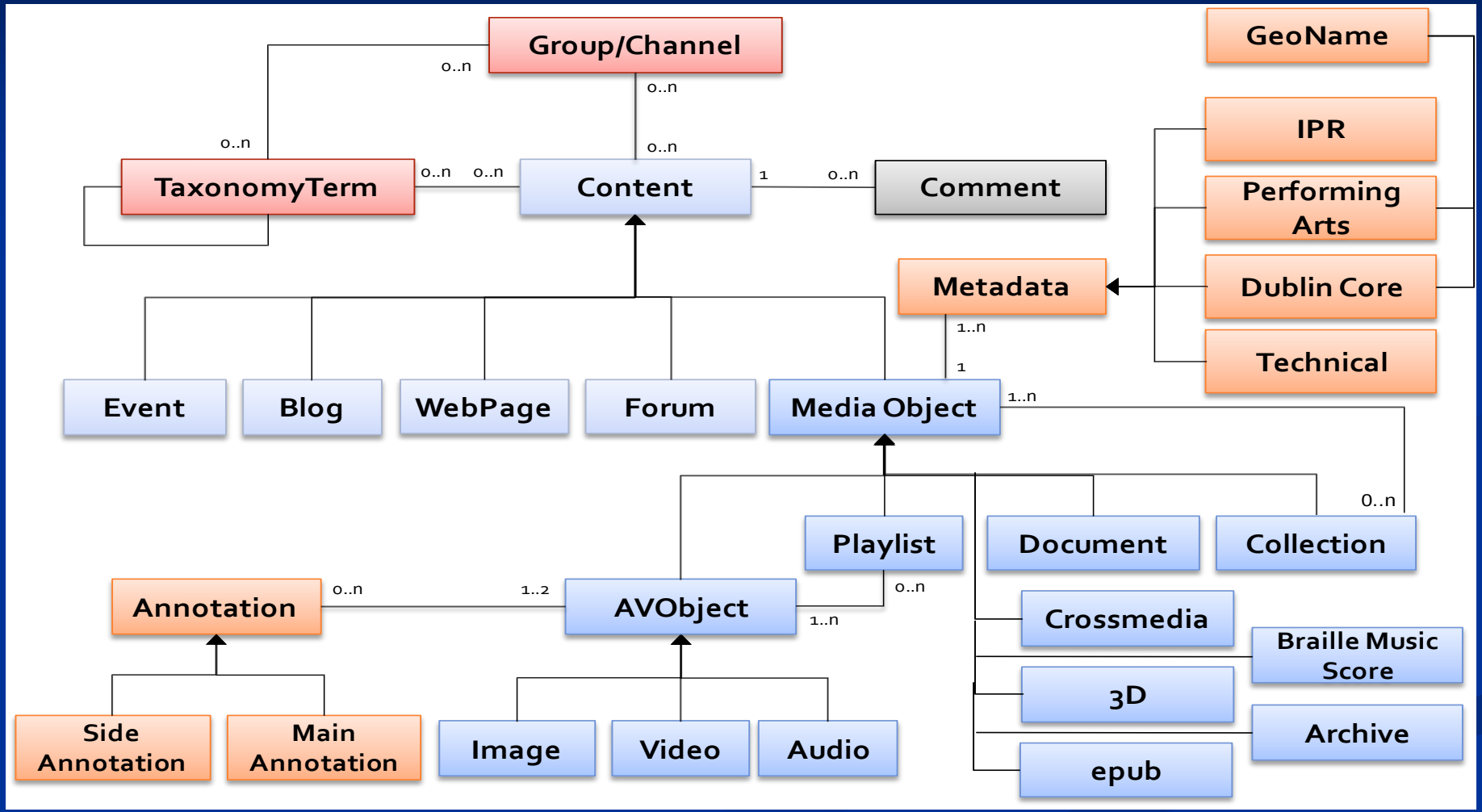
technical

Dynamic

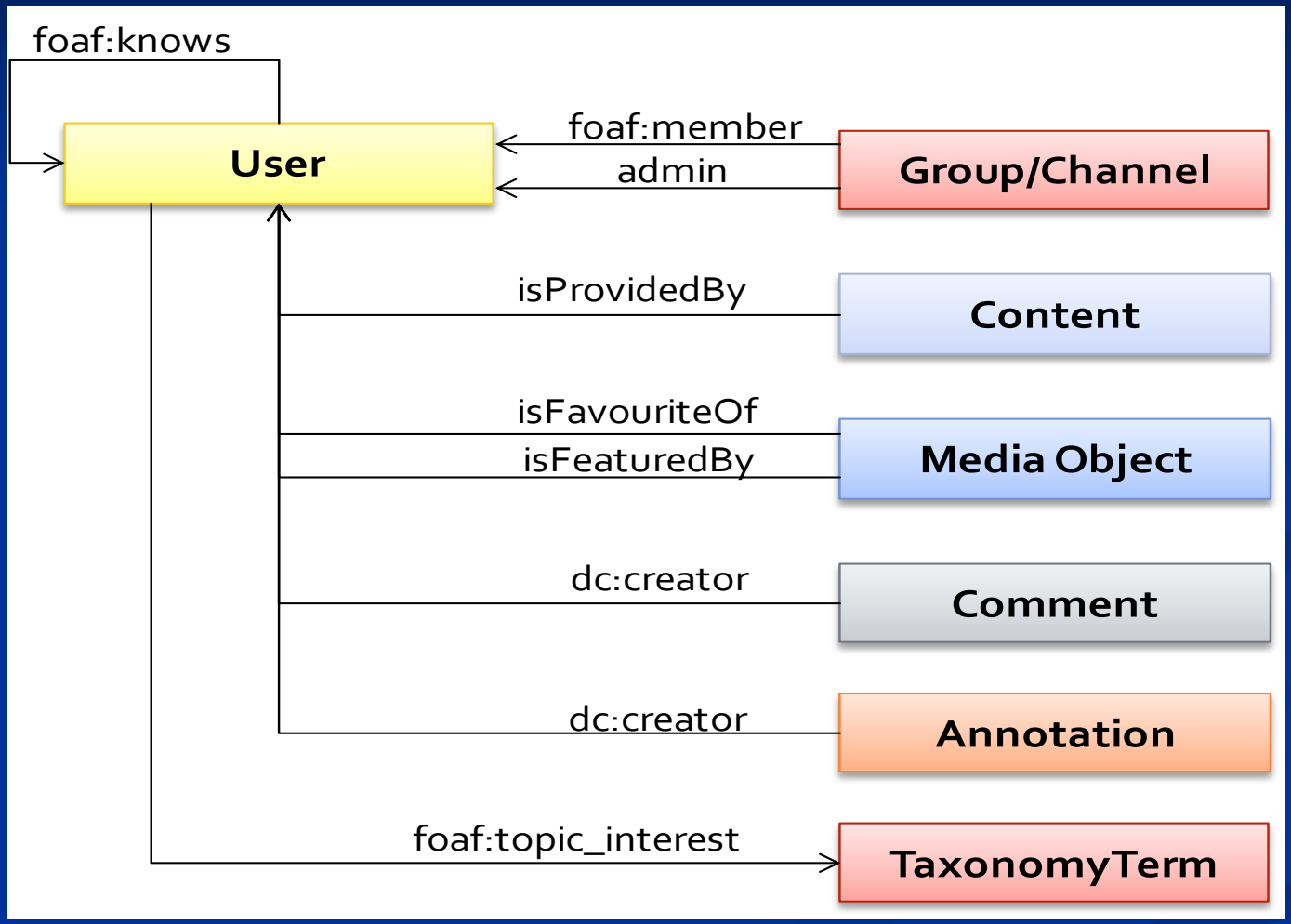
recommend

……

ECLAP Semantic Model 1



ECLAP Semantic Model 2



Indexing

- **Indexing & Search system**
  - Based on Apache Solr
- **Multilingual aspects**
  - Translate the metadata or translate the query?.. both
    - metadata translation
    - Query translation
- **Indexing schema**
  - Dublin Core + DCTerms (multi language)
  - Performing Arts
  - Technical (provider, content type, GPS, IPR, duration, quality, …)
  - Groups associations (multi language)
  - Taxonomy associations (multi language)
  - Comments & multi language tags
  - FullText of the textual digital resources

Indexing

Media Types	DC (ML)	Technical	Performing Arts	Full Text	Tax, Group (ML)	Comments, Tags (ML)	Votes
# of Index Fields*	468	10	23	13	26	13	1
Cross Media: html, MPEG-21, animations, etc.	Y_n	Y	Y	Y	Y_n	Y_m	Y_n
Info text: blog, web pages, events, forum, comments	T	N	N	N	N	Y_m	N
Document: pdf, doc, ePub	Y_n	Y	Y	Y	Y_n	Y_m	Y
Audio, video, image	Y_n	Y	Y	N	Y_n	Y_m	Y_n
Aggregations: play lists, collections, courses, etc.	Y_n	Y	Y	Y/N	Y_n	Y_m	Y_n

* = (# of Fields per Metadata type) × (# of Languages)
ML: Multilingual; DC: Dublin Core; Tax: Taxonomy

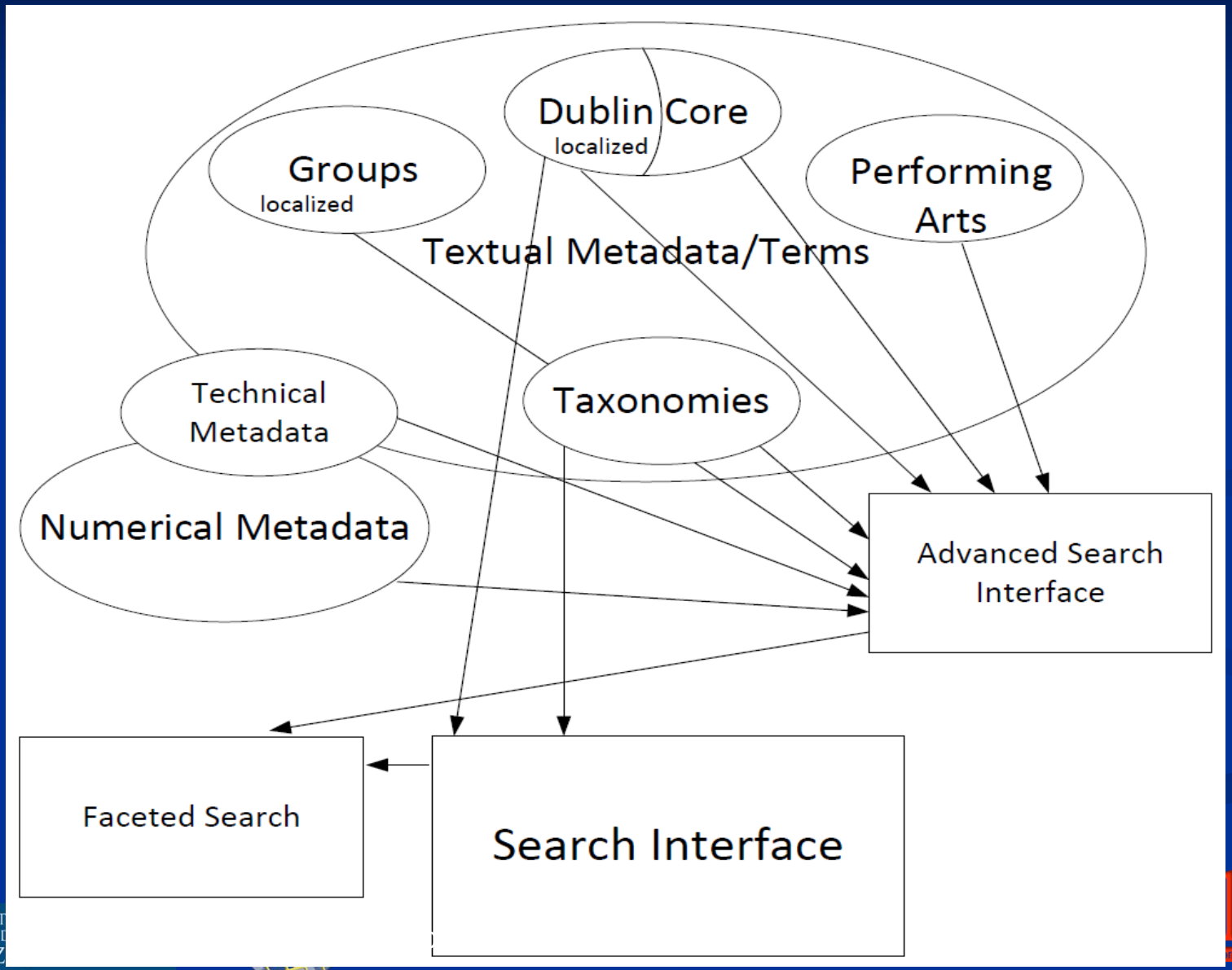
Metadata Schema Indexing

Metadata Type	# fields	Multilingual	Index fields	# fields/item
Performing Arts	23	*N*	23	*n*
Dublin Core	15	*Y*	182	*n*
Dublin Core Terms	22	*Y*	286	*n*
Technical	10	*N*	10	10
Full Text	1	*Y*	13	1
Thematic Groups	1	*Y*	13	20
Taxonomy Terms	1	*Y*	13	231
Pages Comments	13	*N*	13	*n*
Votes	1	*N*	1	1
Total	87	–	554	–

Search Facilities

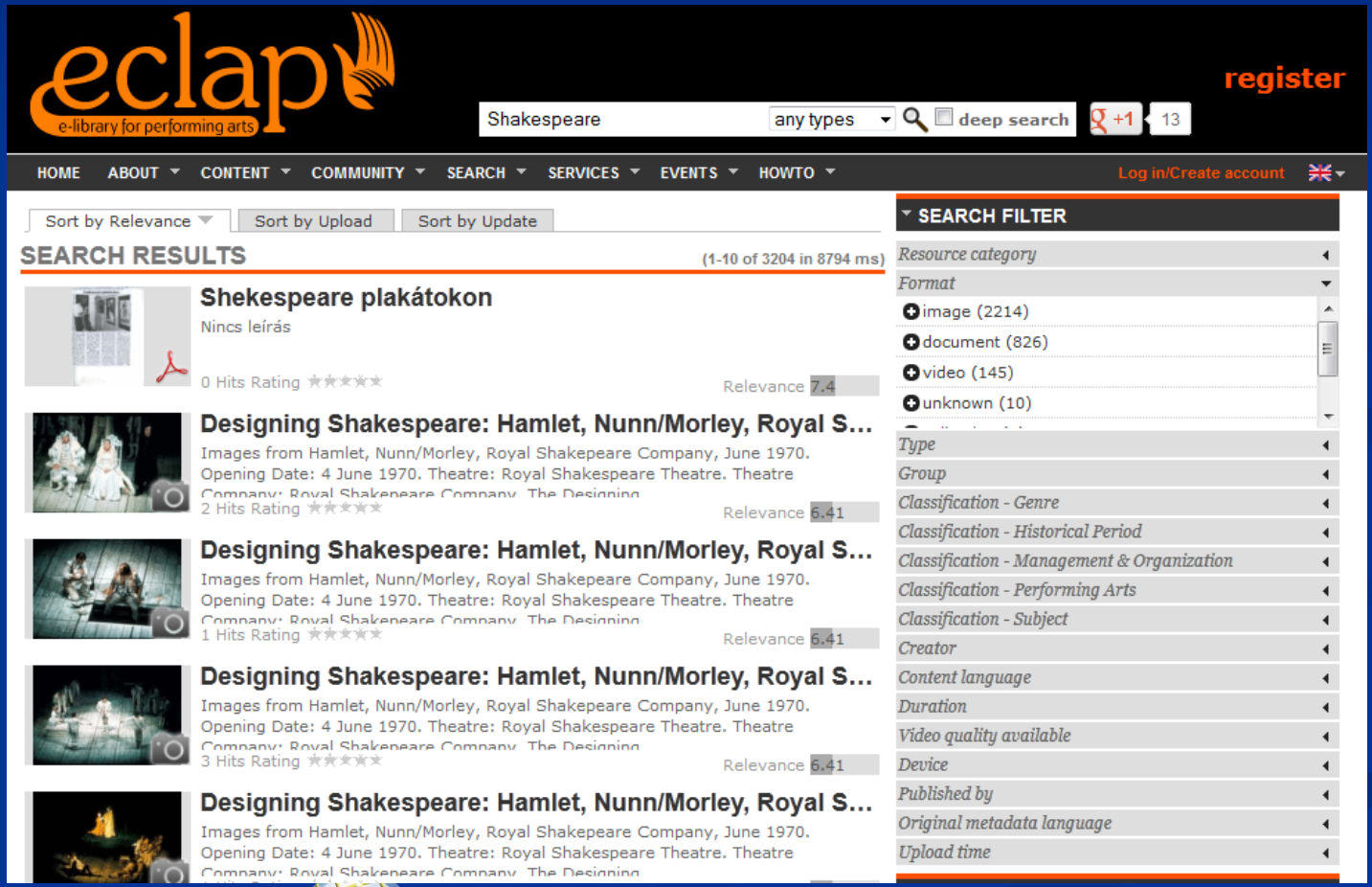
- Full text search
  - Uses the catch all fields to search for keywords in most important fields in all languages (title, description, text, body, subject,…)
- Fuzzy search
  - Allows matching mistyped words
- Deep search
  - Allows searching for partial words
- Faceted Search
- Maximasing Precision and Recall:
  - Relevance & boosting terms

Search Facilities vs Information



Searching

- Faceted search



Weighted Query Model

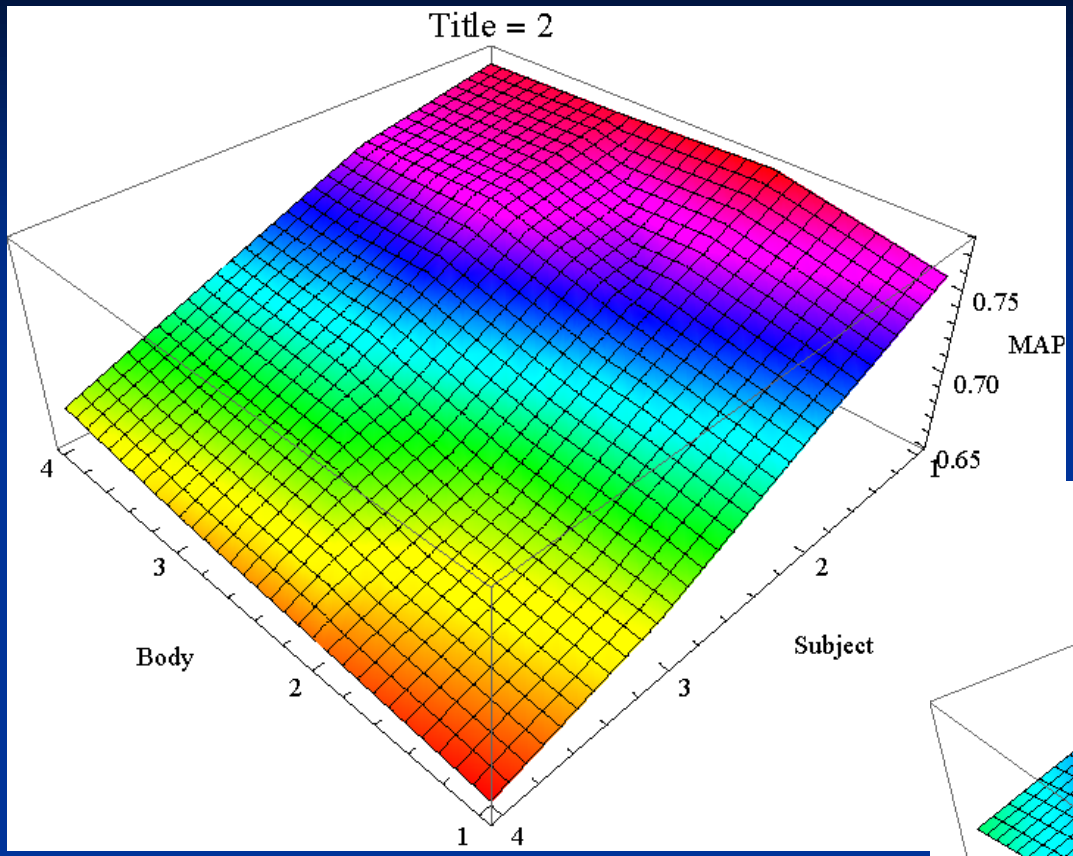
$$b := (title\colon q)^{w_1} \vee (body\colon q)^{w_2} \vee (description\colon q)^{w_3} \vee (subject\colon q)^{w_4} \vee (taxonomy\colon q)^{w_5} \vee (contributor\colon q)^{w_6} \vee (text\colon q)^{w_7}$$

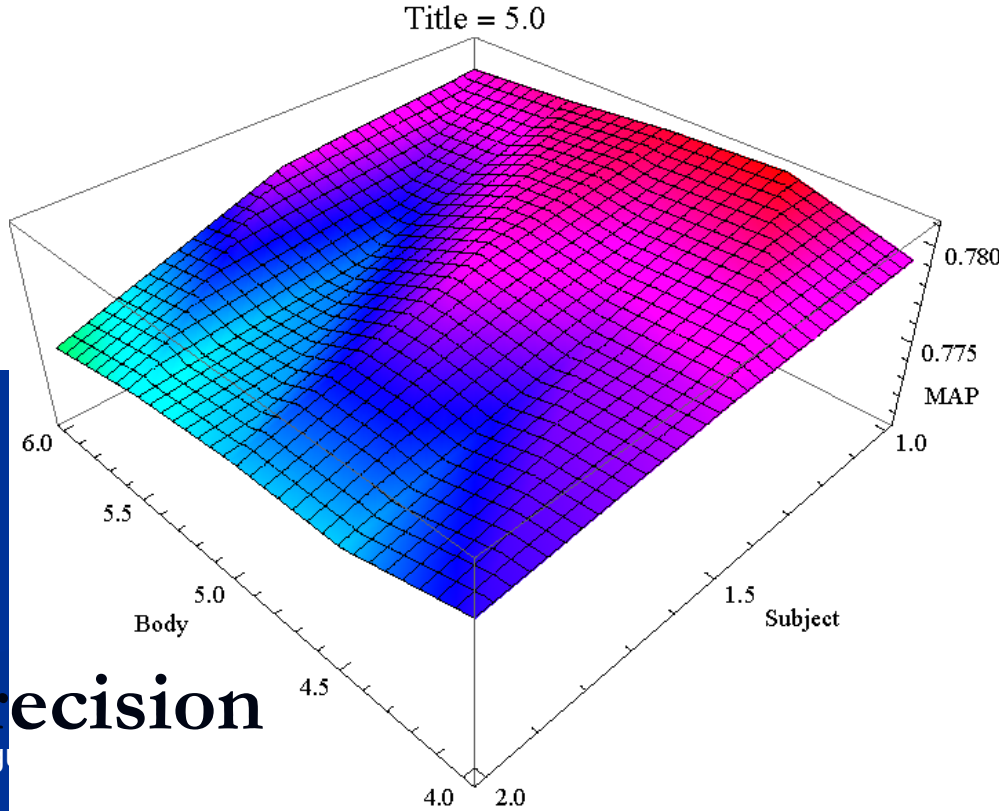
- Where for the "*q*" *query*
  - Weights are boosting fields
    - Title is DC.Title, description DC.Description…..,
    - Body is textual body, subject…,
    - taxonomy the full description of the taxonomy branch

Model Optimization

- Optimization of the Precision&Recall to improve search quality
  - 50 reference queries
- Optimization Methods
  - Simulated Annealing
  - Genetic Algorithms
- 7 parameters

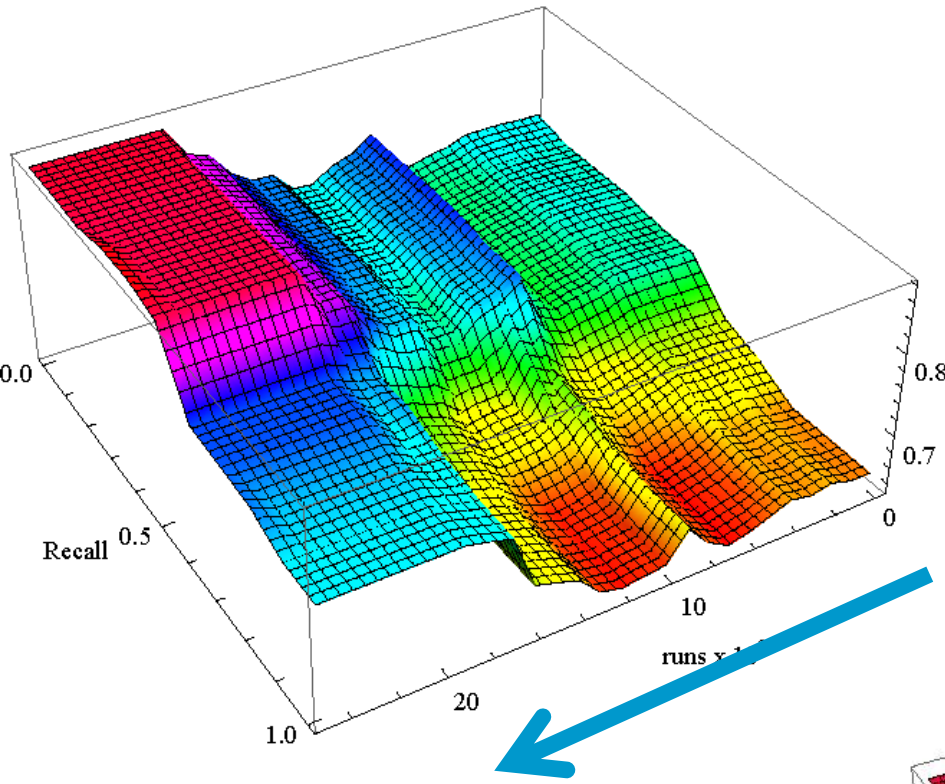
Monte Carlo Analysis



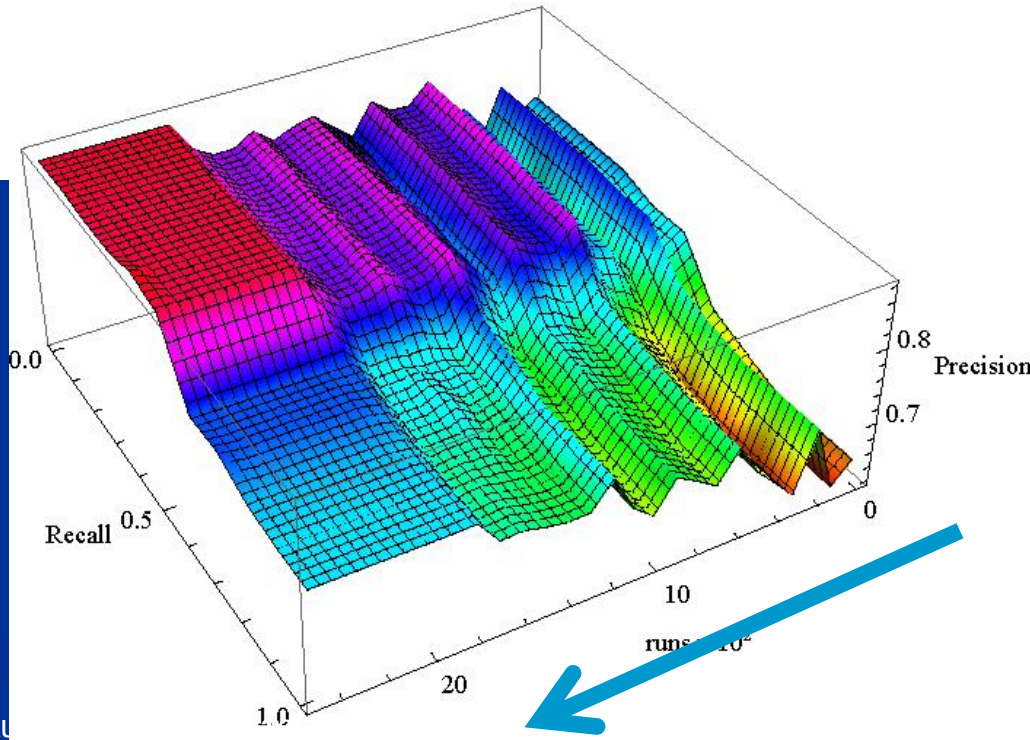


MAP: Mean Average Precision

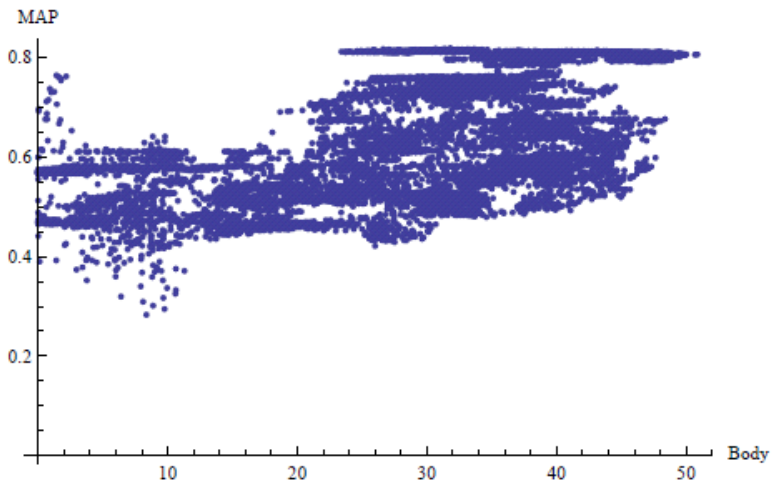
Annealing



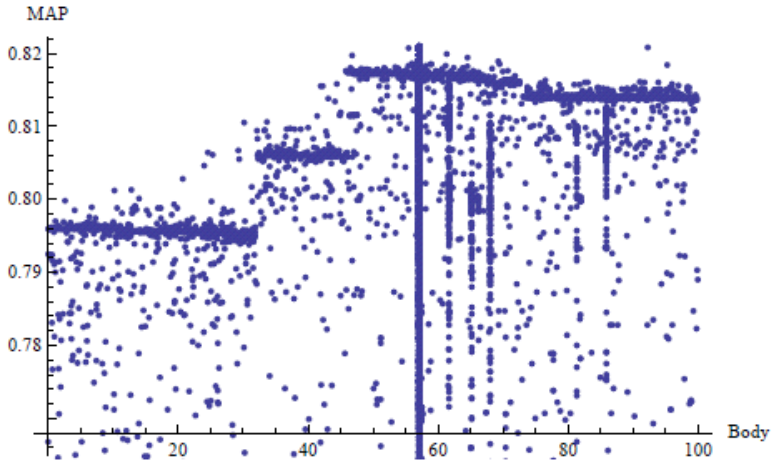
Genetic Alg



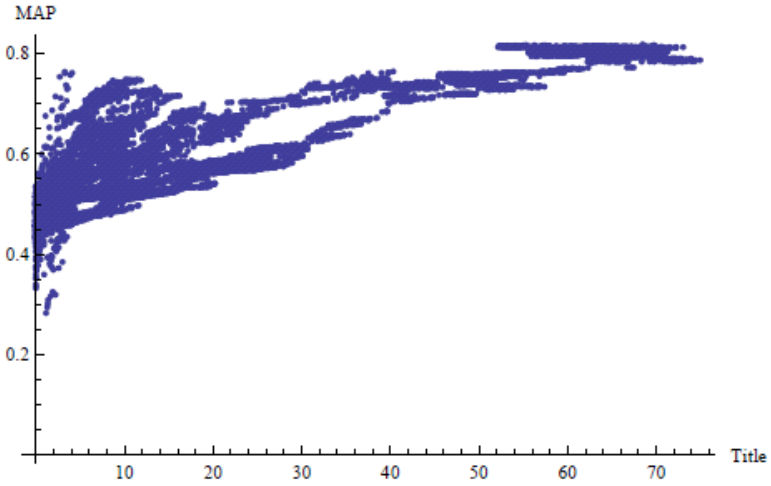
Some weights' Trends



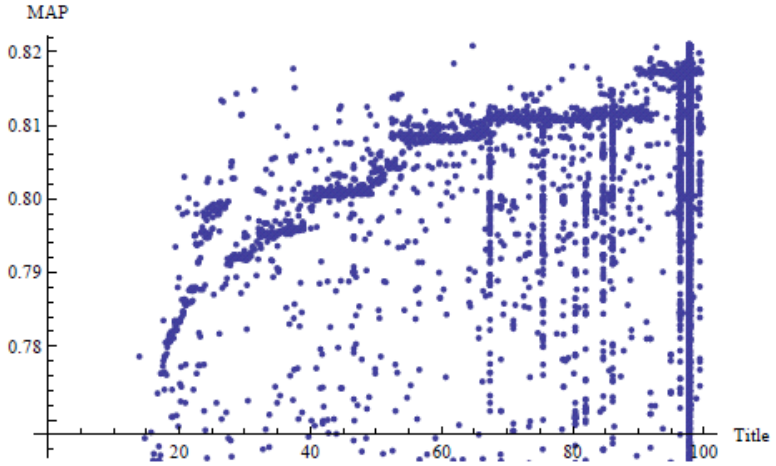
Body vs MAP (Annealing)



Body vs MAP (GA)



Title vs MAP (Annealing)



Title vs MAP (GA)

Comparative Results

Measure	Simulated Annealing	Genetic Algorithm
# of queries	50	50
# of documents retrieved for topic	4312	4319
# of relevant documents for topic	85	85
# of relevant documents retrieved for topic	84	84
MAP	0.8223	0.8210
Geometric MAP	0.7216	0.7169
Precision after retrieving R docs	0.7658	0.7657
Main binary preference measure	0.9886	0.9884
Reciprocal Rank of the first relevant retrieved document	0.8728	0.8747

MAP: Mean Average Precision

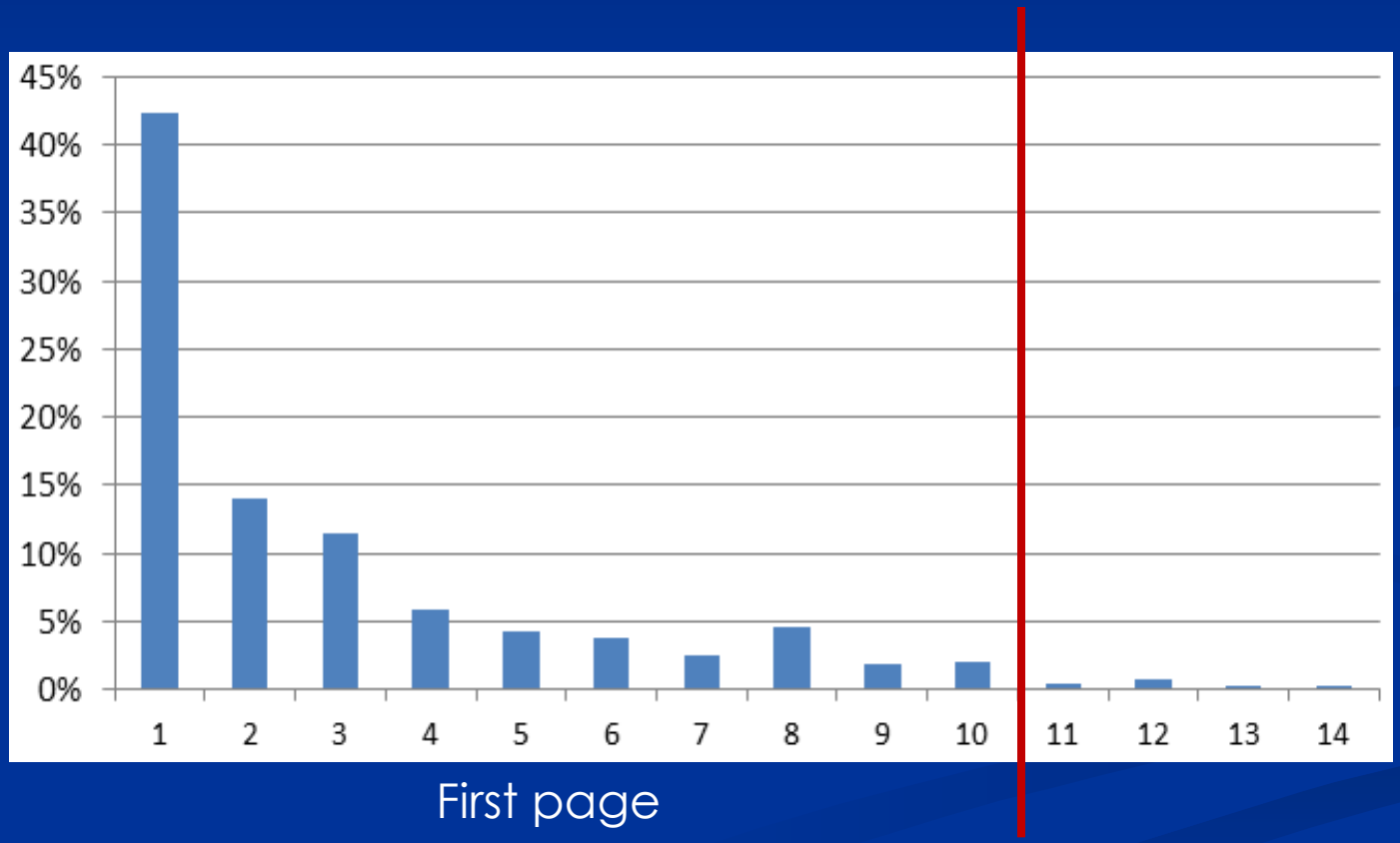
Usage Results

Users	# Full Text Queries	# Faceted Queries	# Last Posted Contents	# Featured Contents	# Popular Contents
Simple Registered	4747	167	34	56	55
Registered as Partners	6665	325	30	91	31
Anonymous	23607	952	1469	533	706
Total	35019	1444	1533	680	792
Clicks after query	17756	589	1150	7448	3407

- Over than 500.000 visits
- 7.29 minutes of permanence on the portal

Assessment of Search Facility

- Distribution of performed clicks



Conclusions

- **indexing solution** for
  - cross media for multilingual metadata and texts
  - Improved Searching & filtering results and thus user experience quality
  - Providing: (**full text, operators)**, **advanced**, **faceted**, etc.
- **Precision and Recall analysis** allowed to tune the search services
  - Simulated Annealing and Genetic Algorithms produced similar results
- **User behavior assessment** has shown that search facility appreciation has been improved wrt to early previous settings, grounded on common sense and classical metadata relevance