Data mining, management and visualization in large scientific corpus

HUI WEI

DRINVENTOR

CARRE

Data collection

Some digital libraries did not supply APIs
We use raw PDF docs as input

Data collection

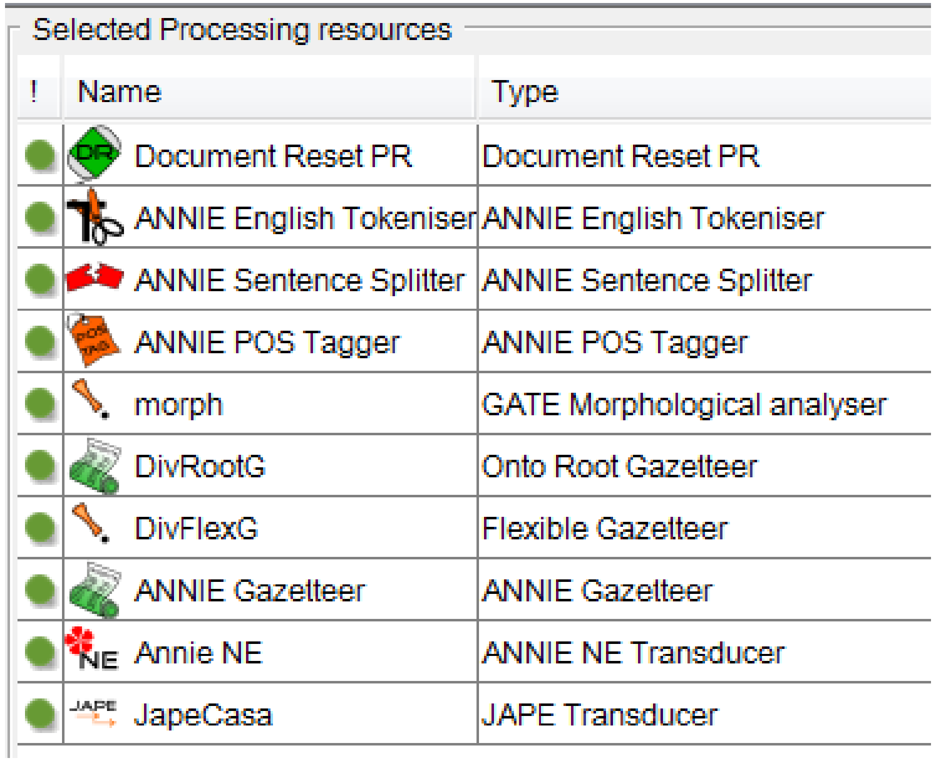
1. to extract basic information of a paper such as authors, title, abstract sentences, doi
2. to extract references
3. to extract standard keywords and their frequency from each paper.

Text mining

1. Use Jape rules to define “Macros” to find important markers, such as”DOI”, “year”, “abstract” tags.

2. Use Annie NE Transducer and Gazetteer look up person names like “author”.

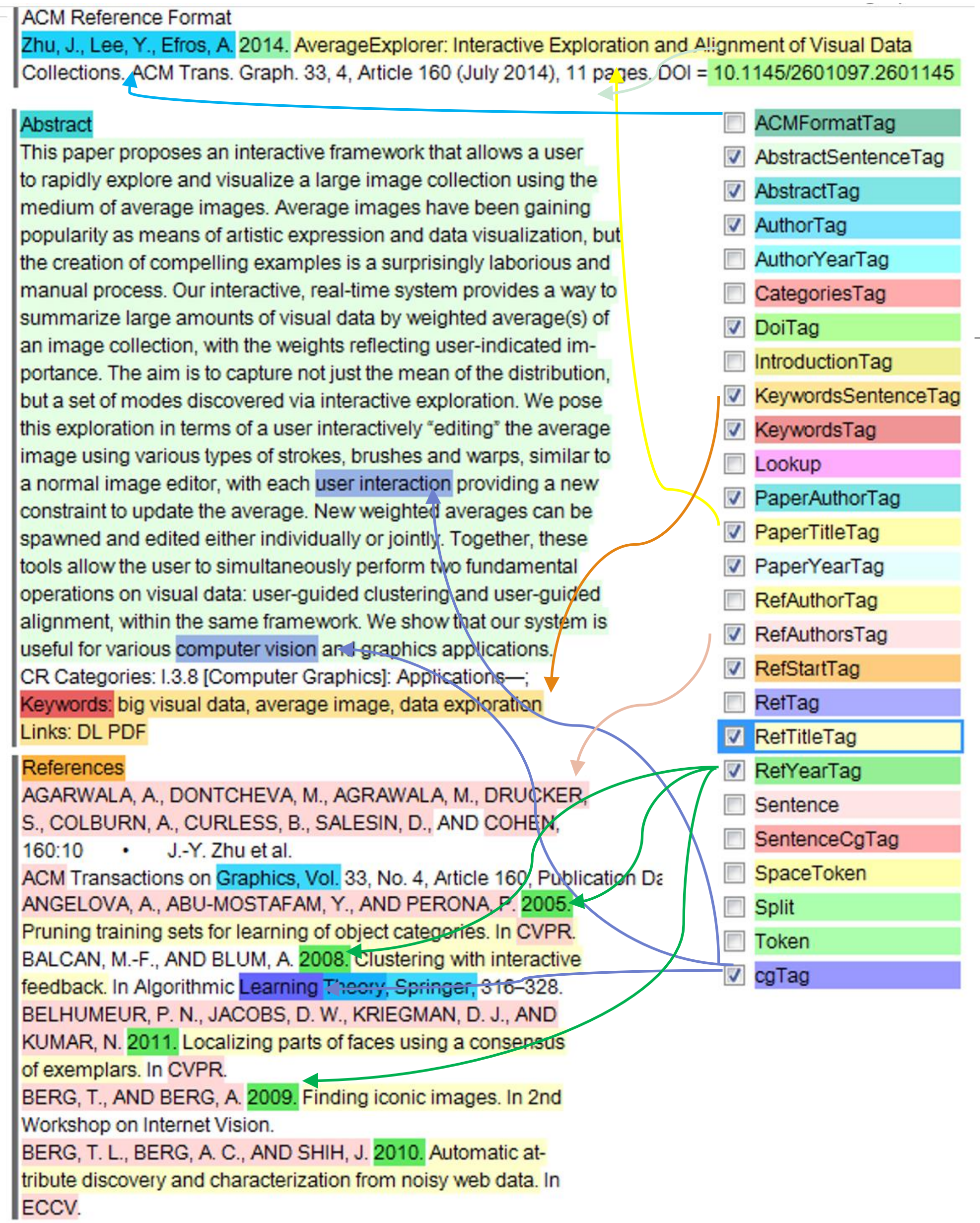
1. Use Gate ontology Gazetteer and Jape rules look up Computer Graphic terms in the content.



Selected Processing resources

!	Name	Type
●	Document Reset PR	Document Reset PR
●	ANNIE English Tokeniser	ANNIE English Tokeniser
●	ANNIE Sentence Splitter	ANNIE Sentence Splitter
●	ANNIE POS Tagger	ANNIE POS Tagger
●	morph	GATE Morphological analyser
●	DivRootG	Onto Root Gazetteer
●	DivFlexG	Flexible Gazetteer
●	ANNIE Gazetteer	ANNIE Gazetteer
●	Annie NE	ANNIE NE Transducer
●	JapeCasa	JAPE Transducer

Text mining



ACM Reference Format

Zhu, J., Lee, Y., Efros, A. 2014. AverageExplorer: Interactive Exploration and Alignment of Visual Data Collections. ACM Trans. Graph. 33, 4, Article 160 (July 2014), 11 pages. DOI = 10.1145/2601097.2601145

Abstract

This paper proposes an interactive framework that allows a user to rapidly explore and visualize a large image collection using the medium of average images. Average images have been gaining popularity as means of artistic expression and data visualization, but the creation of compelling examples is a surprisingly laborious and manual process. Our interactive, real-time system provides a way to summarize large amounts of visual data by weighted average(s) of an image collection, with the weights reflecting user-indicated importance. The aim is to capture not just the mean of the distribution, but a set of modes discovered via interactive exploration. We pose this exploration in terms of a user interactively "editing" the average image using various types of strokes, brushes and warps, similar to a normal image editor, with each user interaction providing a new constraint to update the average. New weighted averages can be spawned and edited either individually or jointly. Together, these tools allow the user to simultaneously perform two fundamental operations on visual data: user-guided clustering and user-guided alignment, within the same framework. We show that our system is useful for various computer vision and graphics applications.

CR Categories: I.3.8 [Computer Graphics]: Applications—;

Keywords: big visual data, average image, data exploration

Links: DL PDF

References

AGARWALA, A., DONTCHEVA, M., AGRAWALA, M., DRUCKER, S., COLBURN, A., CURLESS, B., SALESIN, D., AND COHEN,

160:10 • J.-Y. Zhu et al.

ACM Transactions on Graphics, Vol. 33, No. 4, Article 160, Publication Da

ANGELOVA, A., ABU-MOSTAFAM, Y., AND PERONA, P. 2005. Pruning training sets for learning of object categories. In CVPR.

BALCAN, M.-F., AND BLUM, A. 2008. Clustering with interactive feedback. In Algorithmic Learning Theory, Springer, 316–328.

BELHUMEUR, P. N., JACOBS, D. W., KRIEGMAN, D. J., AND KUMAR, N. 2011. Localizing parts of faces using a consensus of exemplars. In CVPR.

BERG, T., AND BERG, A. 2009. Finding iconic images. In 2nd Workshop on Internet Vision.

BERG, T. L., BERG, A. C., AND SHIH, J. 2010. Automatic attribute discovery and characterization from noisy web data. In ECCV.

☐ ACMFormatTag
☑ AbstractSentenceTag
☑ AbstractTag
☑ AuthorTag
☐ AuthorYearTag
☐ CategoriesTag
☑ DoiTag
☐ IntroductionTag
☑ KeywordsSentenceTag
☑ KeywordsTag
☐ Lookup
☑ PaperAuthorTag
☑ PaperTitleTag
☑ PaperYearTag
☐ RefAuthorTag
☑ RefAuthorsTag
☑ RefStartTag
☐ RefTag
☑ RefTitleTag
☑ RefYearTag
☐ Sentence
☐ SentenceCgTag
☐ SpaceToken
☐ Split
☐ Token
☑ cgTag

Figure 1. Metadata Extraction

Keywords onto

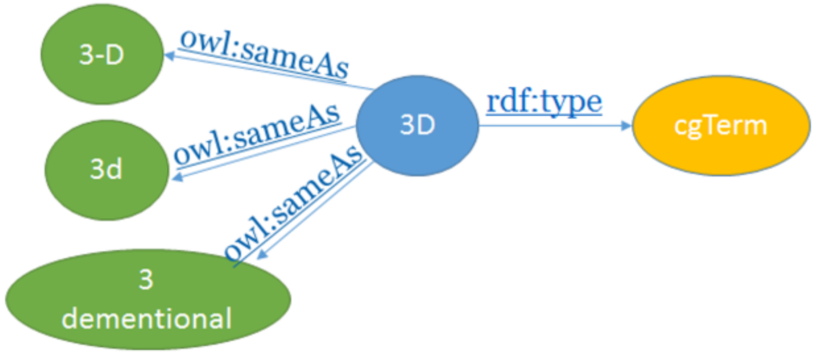
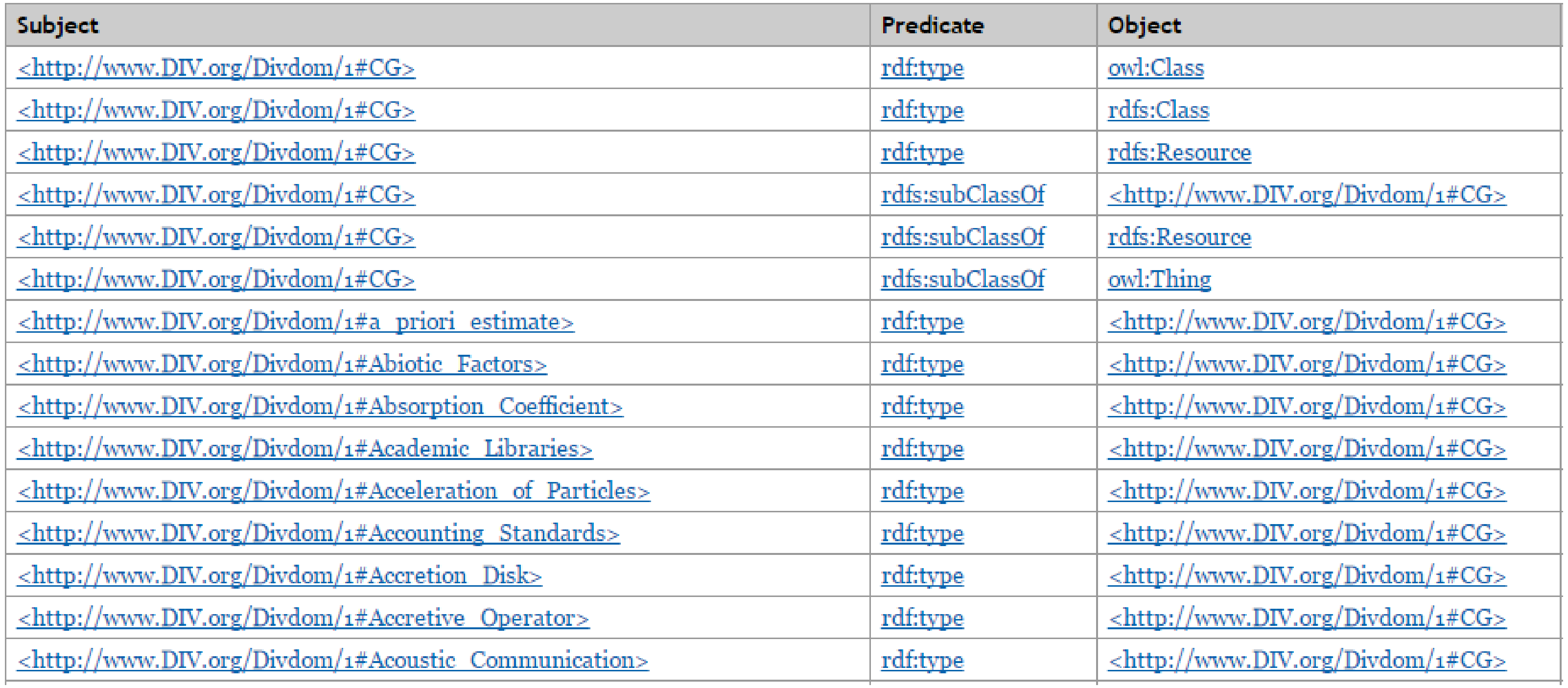
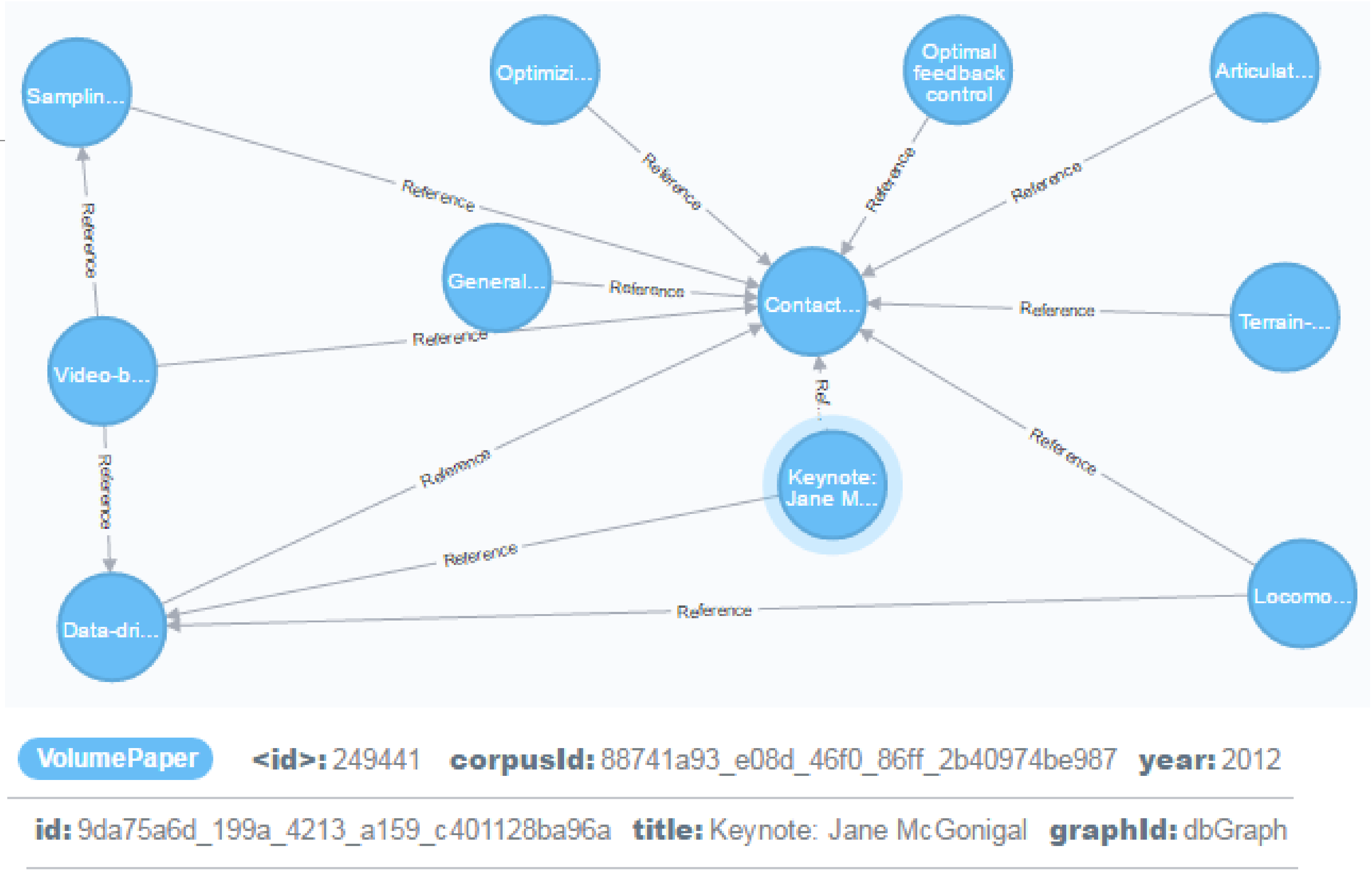


Figure 2. Graph Model



Subject	Predicate	Object
<http://www.DIV.org/Divdom/1#CG>	rdf:type	owl:Class
<http://www.DIV.org/Divdom/1#CG>	rdf:type	rdfs:Class
<http://www.DIV.org/Divdom/1#CG>	rdf:type	rdfs:Resource
<http://www.DIV.org/Divdom/1#CG>	rdfs:subClassOf	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#CG>	rdfs:subClassOf	rdfs:Resource
<http://www.DIV.org/Divdom/1#CG>	rdfs:subClassOf	owl:Thing
<http://www.DIV.org/Divdom/1#a_priori_estimate>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Abiotic_Factors>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Absorption_Coefficient>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Academic_Libraries>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Acceleration_of_Particles>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Accounting_Standards>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Accretion_Disk>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Accretive_Operator>	rdf:type	<http://www.DIV.org/Divdom/1#CG>
<http://www.DIV.org/Divdom/1#Acoustic_Communication>	rdf:type	<http://www.DIV.org/Divdom/1#CG>

Data repositories

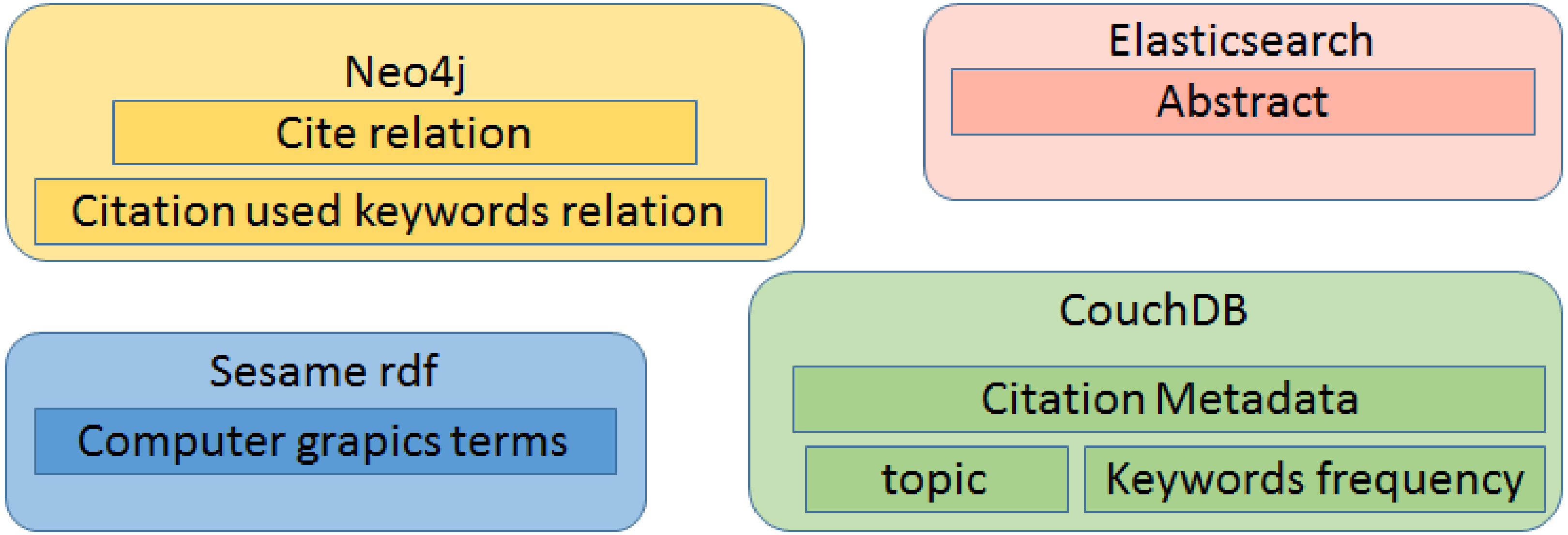


VolumePaper **<id>:** 249441 **corpusId:** 88741a93_e08d_46f0_86ff_2b40974be987 **year:** 2012

id: 9da75a6d_199a_4213_a159_c401128ba96a **title:** Keynote: Jane McGonigal **graphId:** dbGraph

Graph repository

Data repositories



Neo4j

Cite relation

Citation used keywords relation

Elasticsearch

Abstract

Sesame rdf

Computer grapics terms

CouchDB

Citation Metadata

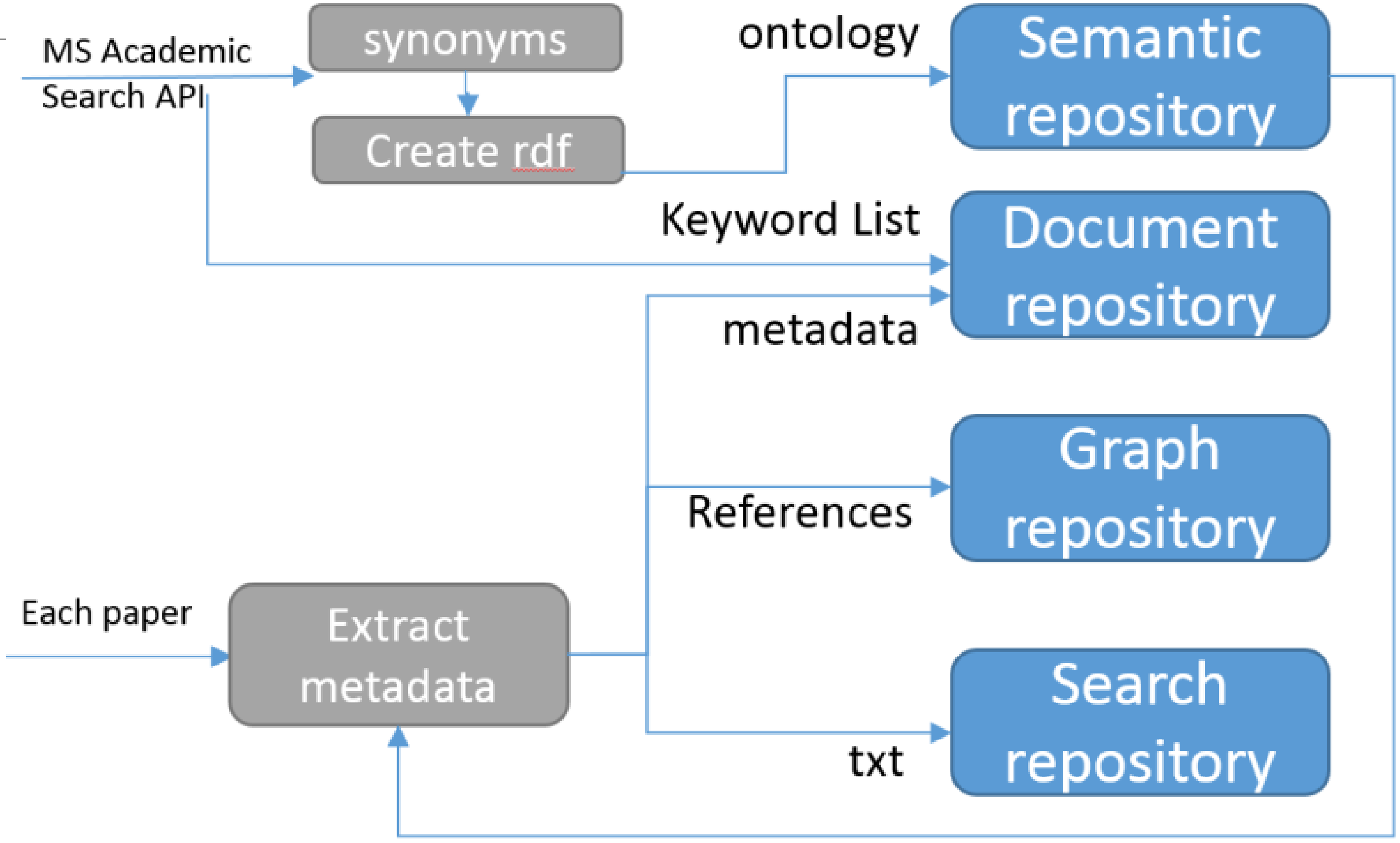
topic

Keywords frequency

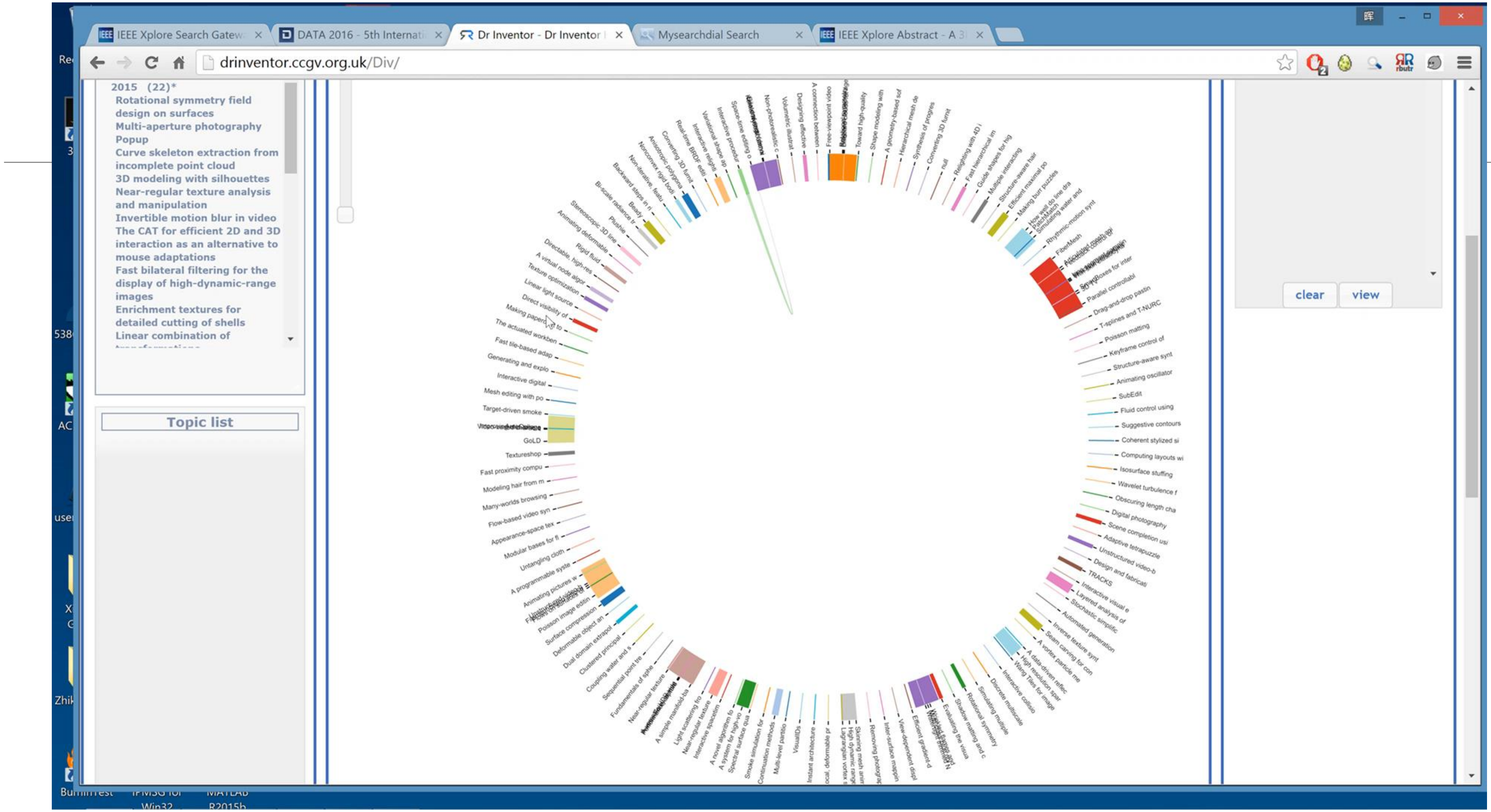
Data is managed in 4 NoSql repositories

Data repositories

Data distribution and system workflow



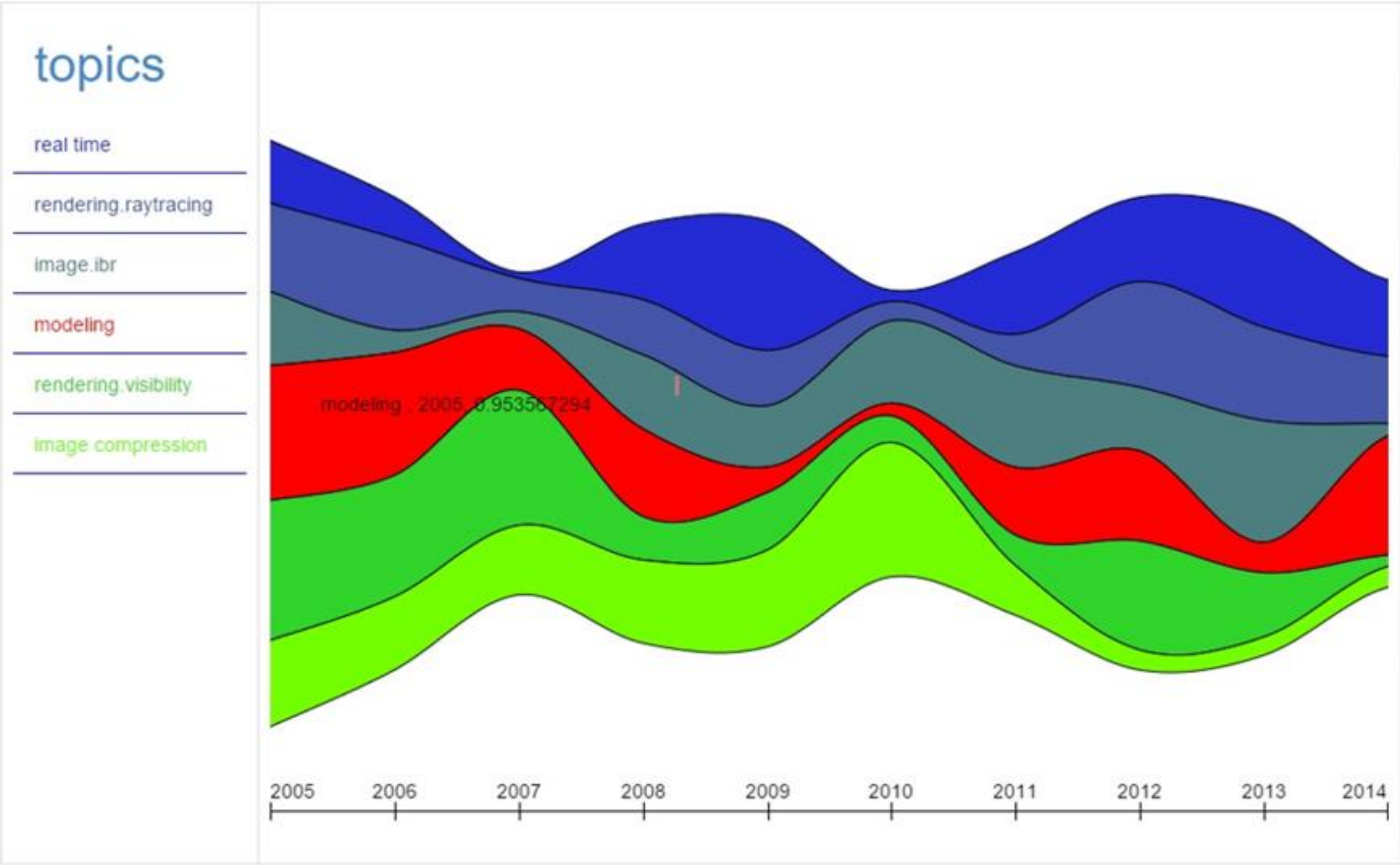
Data visualization



IEEE Xplore Search Gatew
DATA 2016 - 5th Internati
Dr Inventor - Dr Inventor
Mysearchdial Search
IEEE Xplore Abstract - A 3
drinventor.ccgv.org.uk/Div/

2015 (22)*
Rotational symmetry field design on surfaces
Multi-aperture photography
Popup
Curve skeleton extraction from incomplete point cloud
3D modeling with silhouettes
Near-regular texture analysis and manipulation
Invertible motion blur in video
The CAT for efficient 2D and 3D interaction as an alternative to mouse adaptations
Fast bilateral filtering for the display of high-dynamic-range images
Enrichment textures for detailed cutting of shells
Linear combination of transformations

Topic river visualization



Thanks

hui.wei@beds.ac.uk