

University of
South Wales
Prifysgol
De Cymru

JISC

DISTIL

Document Indexing & Semantic Tagging Interface for Libraries

Ceri Binding
Faculty of Computing, Engineering & Science
University of South Wales

<http://hypermedia.research.southwales.ac.uk/>

ceri.binding@southwales.ac.uk



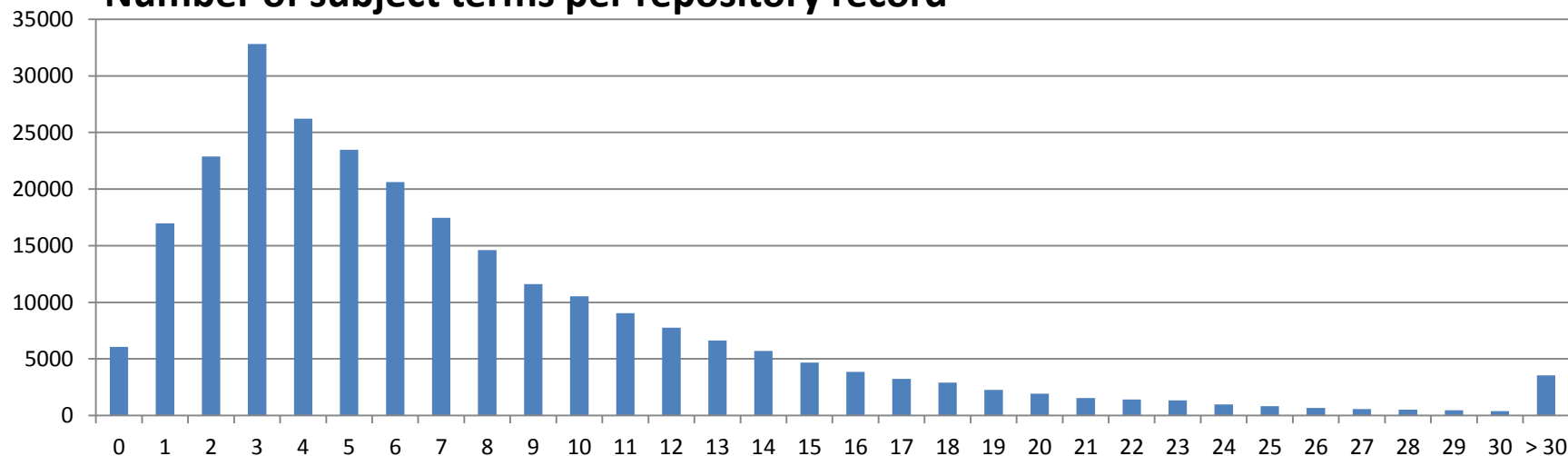
Striking a match...

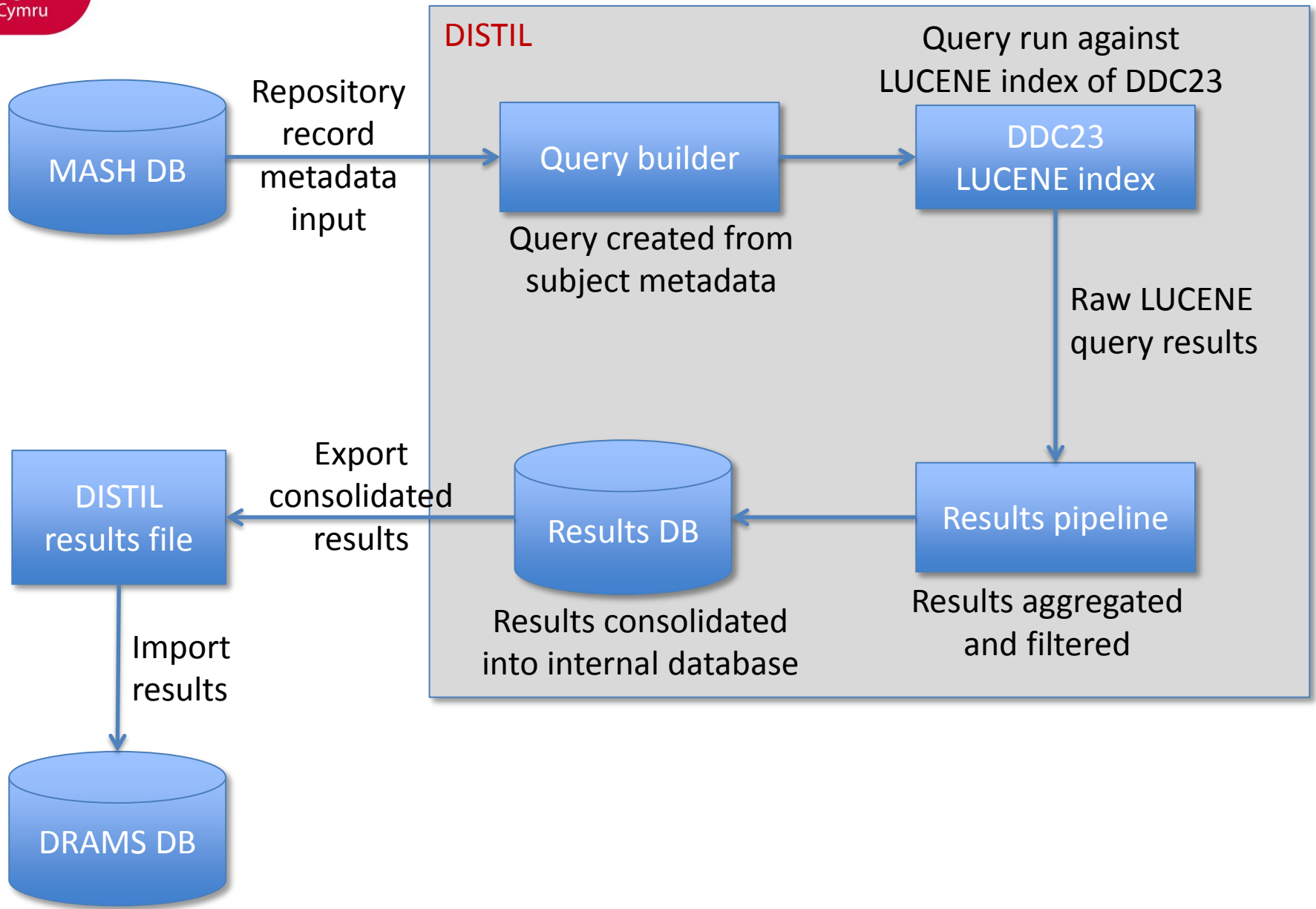
Resource [intute:12345]		
Field type	Field label	Weight
subject	<i>Atmospheric science</i>	1.000
subject	<i>Climatology</i>	1.000
subject	<i>Geoscience</i>	1.000
subject	<i>Meteorology</i>	1.000
title
description
Multiple resource fields of the same type may be present. There are other possible resource field types.		

Match?

DDC Class [551.6]		
Field type	Field label	Weight
label	<i>Climatology and weather</i>	1.000
label	<i>Climate</i>	1.000
label	<i>Climatology</i>	1.000
label	<i>Weather</i>	1.000
Multiple terminology fields of the same type may be present.		

Number of subject terms per repository record

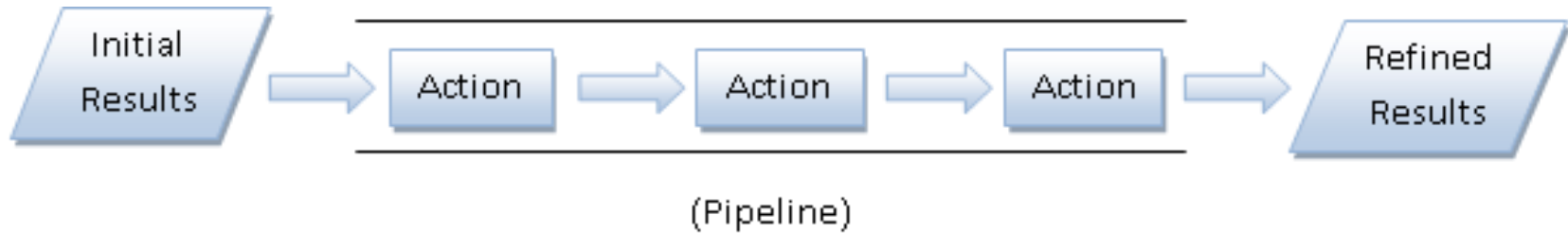




Creating LUCENE queries from metadata

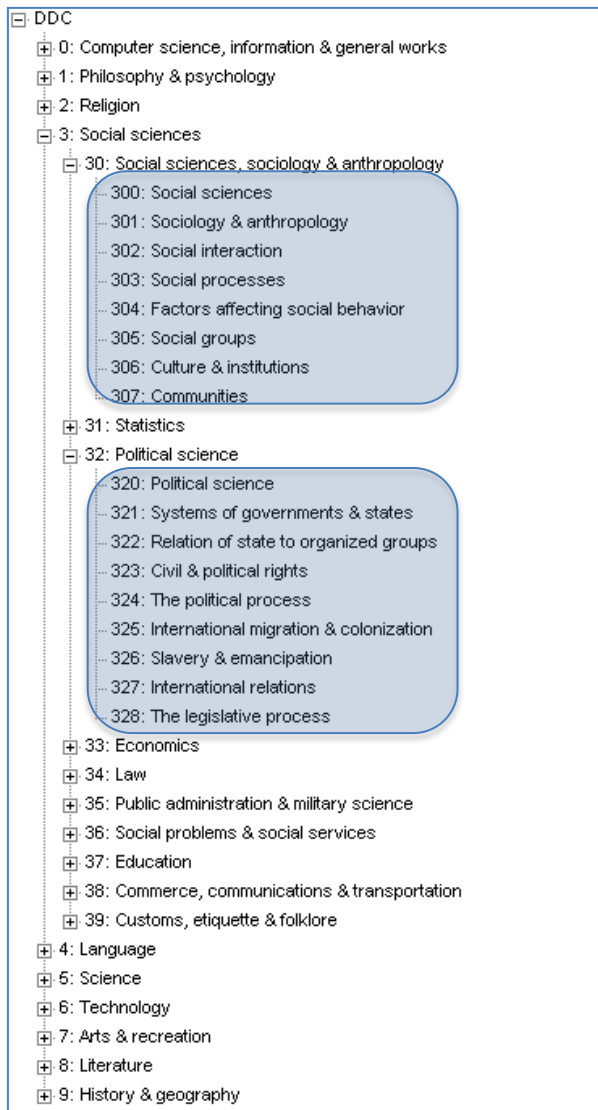
- Metadata elements
 - **Subjects**, titles, descriptions
- Term stemming - Porter stemming algorithm
 - **educate/educating/ educational /education** → label:**educ**
- Stop words – remove common words
 - “Safety **and** mitigation” →
label:"Safety and Mitigation" (+label:safeti +label:mitig)
- Phrase parsing
 - Exact match, or stopped and stemmed words present in any order
 - “Eye Diseases” → label:"Eye Diseases" (+label:ey +label:diseas)
- Character escaping (LUCENE query reserved characters)
 - “Art, Music **&** Museums” →
label:"Art, Music \& Museums" (+label:art +label:music +label:museum)
- US/UK spelling differences – supplement query terms
 - “theatres” → (label:theatre label:theater)

DISTIL results pipeline



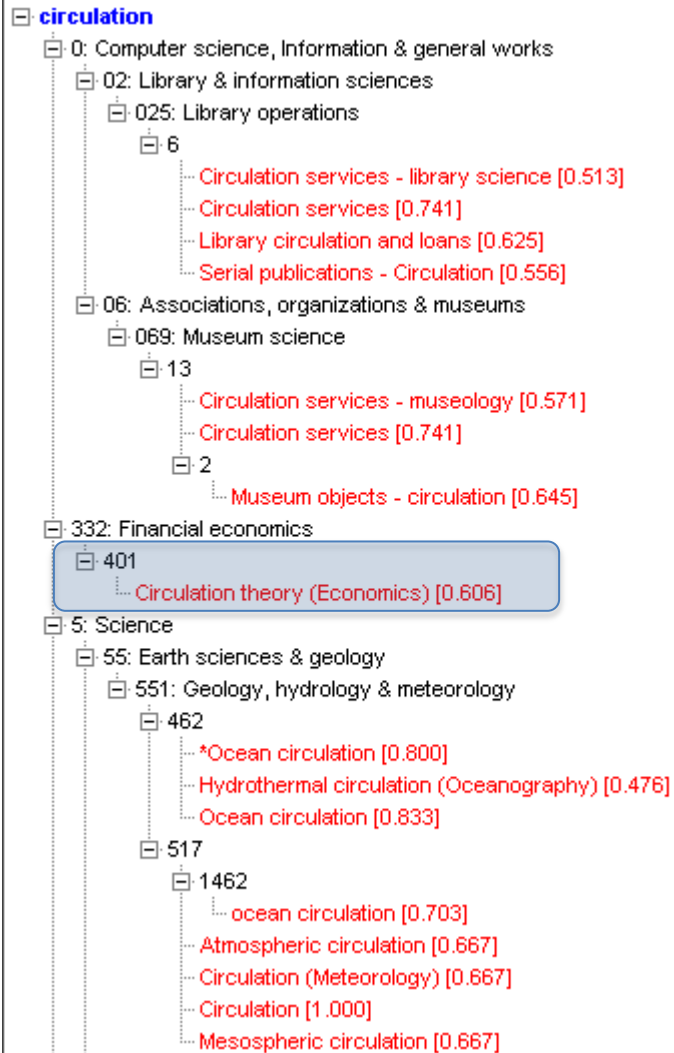
- Results Pipeline Actions
 - DdcSummaryLevelMinAction
 - DdcRemoveOutliersAction
 - DdcUseAbridgedIdAction
 - DdcRuleOfThreeAction
 - DdcAddSumDescendantScoreAction
 - DdcRemoveDescendantsAction
 - DdcAddDominantSummaryScores
 - NormalizeValuesAction
 - SortRowsAction
 - LimitRowsAction

DdcSummaryLevelMinAction



- (DDC Intro, p.37, section 13.3):
“The classifier should never reduce the notation to less than the most specific three-digit number
- DDC Summaries are the top 3 hierarchical levels
- Top 2 levels are for structure only – indexing should use as minimum 3rd level (3 digits)
- Any initial matches on level 1 or 2 are deleted from results

DdcRemoveOutliersAction



- Outliers are isolated classes having no other ancestor or descendant matches on *any* of the query terms
- Outliers are removed from the results

DdcUseAbridgedIdAction

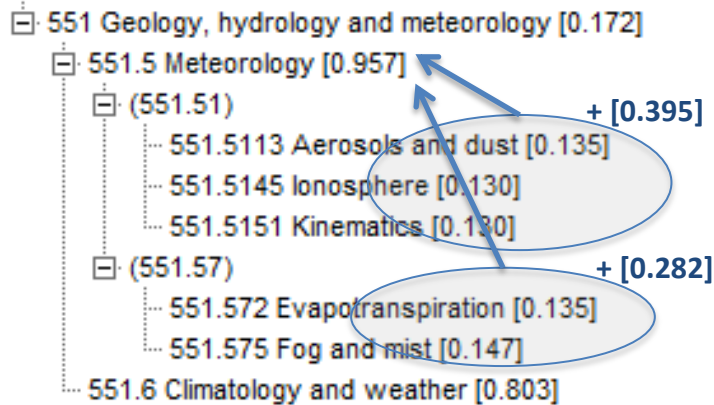
- Upward score aggregation, from ‘close’ classification to ‘broad’ classification.
- E.g. a work on French cooking is classed closely at 641.5944 (641.59 Cooking by place + 44 France from Table 2), or broadly at 641.5 (Cooking).
- Broad classification means that *“the work is placed in a broad class by use of notation that has been logically abridged”*.
- The broad class (a.k.a. abridged number) is not necessarily the direct parent class

DdcRuleOfThreeAction

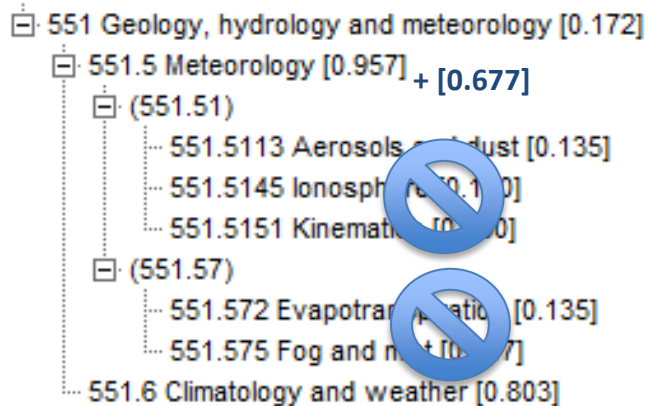


- (DDC Intro, p.8 section 5.7D):
“Class a work on three or more subjects that are all subdivisions of a broader subject in the first higher number that includes them all”
- Any 3 (or more) matching classes with a common parent are replaced with that parent.
- If the parent is not already present in the results it is added.
- The sum scores of the children are added to the parent and the children removed from the results.

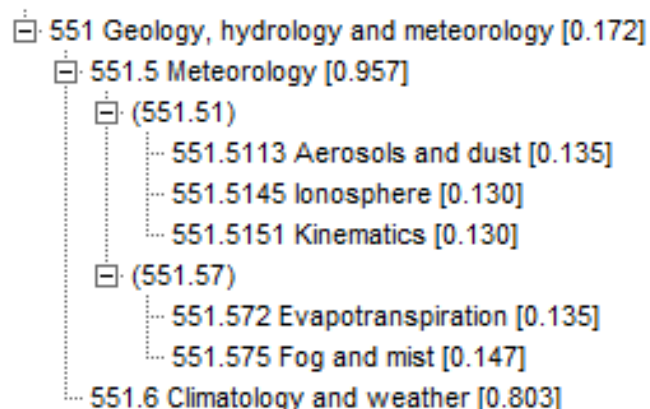
DdcAddSumDescendantScoreAction + DdcRemoveDescendantsAction



- Upward score aggregation - a matching class can inherit the aggregated scores of any descendant matches also present
- The sum of scores for descendant matches are added to the ancestor class
- Classes that have contributed to a parent class score are then removed from the results



DdcAddDominantSummaryScores



All results: overall sum = 38.529

Level 1 ("5") sum = 12.376

Level 2 ("55") sum = 8.443

Level 3 ("551") sum = 2.609

For each result:

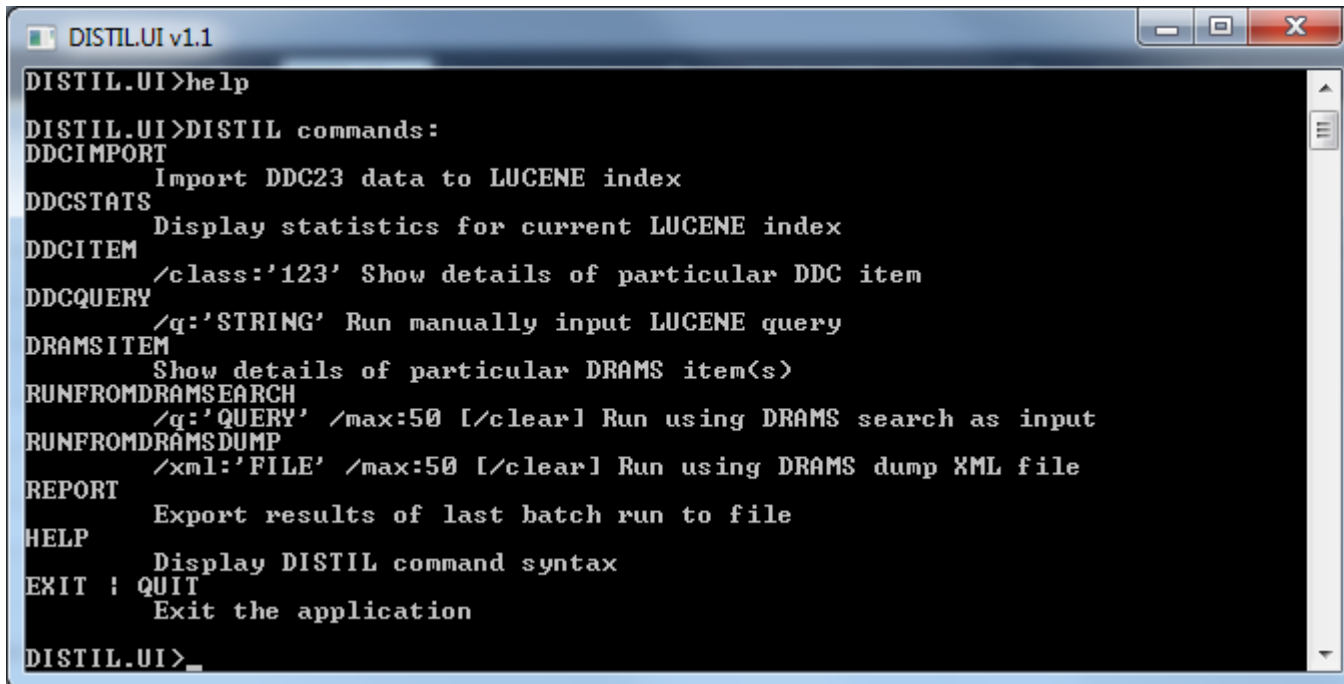
New score = ((score + L1 sum + L2 sum + L3 sum) / overall sum)

- Experimental ‘top down aggregation’
- Boosts the scores of classes originating from the ‘dominant’ summaries in the results
- Helps to promote results from particularly strong subject areas
- Makes a good result better; but sometimes makes a bad result worse!

Others...

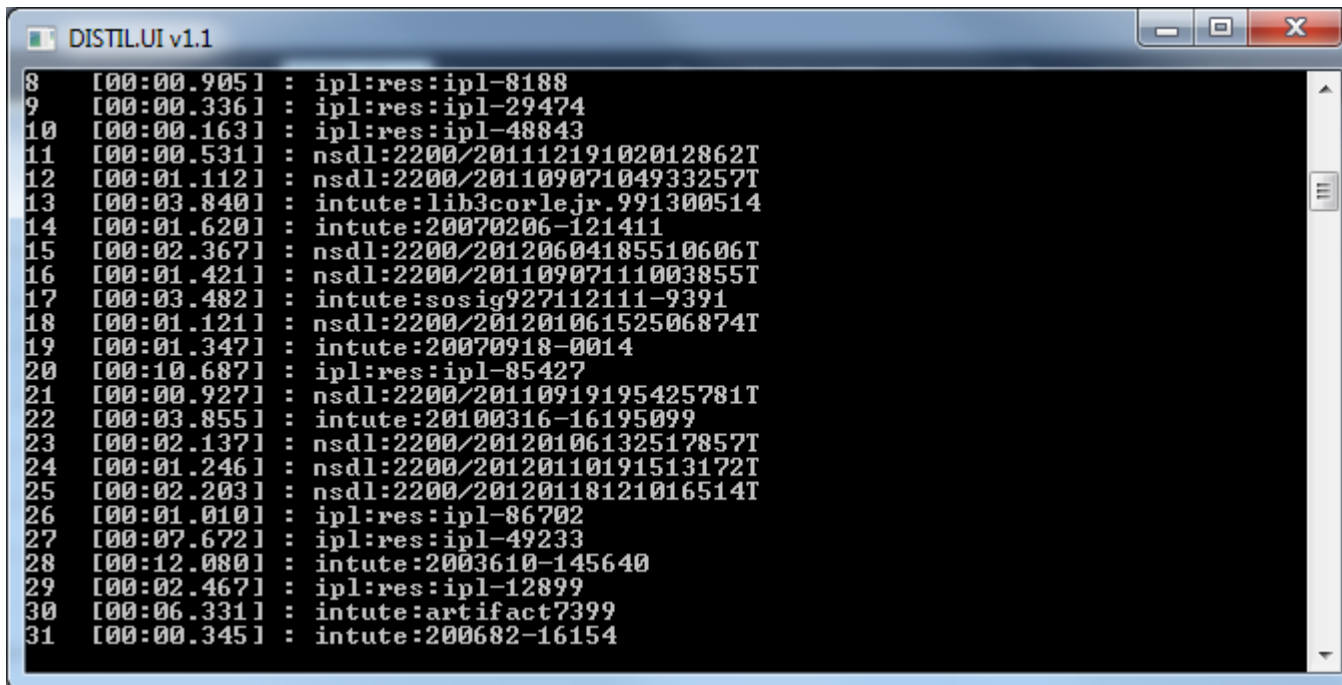
- **DdcNormalizeValuesAction**
 - Scores are normalized to give relative ranked scores in the range 0..1
 - $$x_{new} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$
- **SortRowsAction**
 - Results are sorted in ascending score order
- **LimitRowsAction**
 - Results are limited to the top 10 results

DISTIL.UI application - commands

A screenshot of a terminal window titled "DISTIL.UI v1.1". The window has a standard Windows-style title bar with minimize, maximize, and close buttons. The terminal content shows a list of commands and their descriptions. The user has entered "help" and the application has responded with a list of commands. The list includes: DDCIMPORT (Import DDC23 data to LUCENE index), DDCSTATS (Display statistics for current LUCENE index), DDCITEM (/class:'123' Show details of particular DDC item), DDCQUERY (/q:'STRING' Run manually input LUCENE query), DRAMSITEM (Show details of particular DRAMS item(s)), RUNFROMDRAMSEARCH (/q:'QUERY' /max:50 [/clear] Run using DRAMS search as input), RUNFROMDRAMS_DUMP (/xml:'FILE' /max:50 [/clear] Run using DRAMS dump XML file), REPORT (Export results of last batch run to file), HELP (Display DISTIL command syntax), and EXIT : QUIT (Exit the application). The prompt "DISTIL.UI>_" is visible at the bottom.

```
DISTIL.UI v1.1
DISTIL.UI>help
DISTIL.UI>DISTIL commands:
DDCIMPORT
    Import DDC23 data to LUCENE index
DDCSTATS
    Display statistics for current LUCENE index
DDCITEM
    /class:'123' Show details of particular DDC item
DDCQUERY
    /q:'STRING' Run manually input LUCENE query
DRAMSITEM
    Show details of particular DRAMS item(s)
RUNFROMDRAMSEARCH
    /q:'QUERY' /max:50 [/clear] Run using DRAMS search as input
RUNFROMDRAMS_DUMP
    /xml:'FILE' /max:50 [/clear] Run using DRAMS dump XML file
REPORT
    Export results of last batch run to file
HELP
    Display DISTIL command syntax
EXIT : QUIT
    Exit the application
DISTIL.UI>_
```

DISTIL.UI batch processing of repository records



```
DISTIL.UI v1.1
8 [00:00.905] : ipl:res:ipl-8188
9 [00:00.336] : ipl:res:ipl-29474
10 [00:00.163] : ipl:res:ipl-48843
11 [00:00.531] : nsdl:2200/20111219102012862T
12 [00:01.112] : nsdl:2200/20110907104933257T
13 [00:03.840] : intute:lib3corlejr.991300514
14 [00:01.620] : intute:20070206-121411
15 [00:02.367] : nsdl:2200/20120604185510606T
16 [00:01.421] : nsdl:2200/20110907111003855T
17 [00:03.482] : intute:sosig927112111-9391
18 [00:01.121] : nsdl:2200/20120106152506874T
19 [00:01.347] : intute:20070918-0014
20 [00:10.687] : ipl:res:ipl-85427
21 [00:00.927] : nsdl:2200/20110919195425781T
22 [00:03.855] : intute:20100316-16195099
23 [00:02.137] : nsdl:2200/20120106132517857T
24 [00:01.246] : nsdl:2200/20120110191513172T
25 [00:02.203] : nsdl:2200/20120118121016514T
26 [00:01.010] : ipl:res:ipl-86702
27 [00:07.672] : ipl:res:ipl-49233
28 [00:12.080] : intute:2003610-145640
29 [00:02.467] : ipl:res:ipl-12899
30 [00:06.331] : intute:artifact7399
31 [00:00.345] : intute:200682-16154
```

- Processing bulk XML downloads from DRAMS SOLR API
- Runs slowly but surely... (approx. 1000 records in 30 minutes)
- Initially 100,000 repository records processed & results sent to DREXEL

DISTIL.UI outputs – tabular data

- Repository
- Record ID
- DDC class
- Score
- DDC Label

lib	rid	ddc	score	label
intute	2001689	617	1	Surgery, regional medicine, dentistry, ophthalmology, otology, audiology
intute	2001689	618.921	0.109056	Regional medicine, ophthalmology, otology, audiology
intute	2001689	573.88	0.080421	Eyes
intute	2001689	612.84	0.024369	Eyes
intute	2003428-2	343	1	Military, defense, public property, public finance, tax, commerce (trade), industrial
intute	2003428-2	336	0.902761	Public finance
intute	2003428-2	613	0.819975	Personal health and safety
intute	2003428-2	371	0.766684	Schools and their activities; special education
intute	2003428-2	372	0.755531	Primary education (Elementary education)
intute	2003428-2	378	0.705374	Higher education (Tertiary education)
intute	2003428-2	370	0.648131	Education
intute	2003428-2	344	0.647601	Labor, social service, education, cultural law
intute	2003428-2	379	0.612226	Public policy issues in education
intute	2003428-2	331.252	0.51913	Pensions
intute	2003610-1	973	1	United States
intute	2003610-1	809	0.41955	History, description, critical appraisal of more than two literatures
intute	2003610-1	970	0.363838	History of North America
intute	2003610-1	979	0.298301	Great Basin and Pacific Slope region of United States
intute	2003610-1	551.2	0.272352	Volcanoes, earthquakes, thermal waters and gases
intute	2003610-1	271	0.239134	Religious congregations and orders in church history
intute	2003610-1	348.73	0.229985	Federal laws, regulations, cases of the United States
intute	2003610-1	347.73	0.22878	Civil procedure and courts of the United States
intute	2003610-1	936	0.226373	Europe north and west of Italian Peninsula to ca. 499
intute	2003610-1	912	0.226131	Graphic representations of surface of earth and of extraterrestrial worlds
intute	2004622-9	661	1	Technology of industrial chemicals
intute	2004622-9	660	0.635892	Chemical engineering and related technologies
intute	2004622-9	547	0.566724	Organic chemistry
intute	2004622-9	572	0.380721	Biochemistry

DISTIL.UI outputs – textual report

- ID
- Metadata
- Query
- Pipeline
- Matches

```

ID: intute|2001689 [http://www.cgeye.org/]
Metadata:
subject : Eye [1.000]
subject : Ophthalmology [1.000]
subject : Eye Diseases [1.000]
url : https://www.cgeye.org/ [0.000]
description : A digital image library containing images of eye conditions including fluorescein a
diabetic retinopathy screening and treatment procedures, the external eye, anterior eye, glaucoma
on the Web by Dr Peter Scanlon, an associate specialist in ophthalmology at Cheltenham General Ho
title : Eye library [0.000]

Query:
label:ey label:ophthalmolog label:"Eye Diseases" (+label:ey +label:diseas)
Pipeline:
DISTIL.Pipeline.DdcSummaryLevelMinAction
DISTIL.Pipeline.DdcRemoveOutliersAction
DISTIL.Pipeline.DdcRuleOfThreeAction
DISTIL.Pipeline.DdcAddSumDescendantScoreAction
DISTIL.Pipeline.DdcRemoveDescendantsAction
DISTIL.Pipeline.NormalizeValuesAction(colName: 'score')
DISTIL.Pipeline.SortRowsAction(rowSort: 'score DESC')
DISTIL.Pipeline.LimitRowsAction(maxRows: 10)

Matches:
id          score label
617         1.000 Surgery, regional medicine, dentistry, ophthalmology, otology, audiology
618.92097   0.109 Regional medicine, ophthalmology, otology, audiology
573.88      0.080 Eyes
612.84      0.024 Eyes
-----
ID: intute|2003428-20109g [http://www.ifs.org.uk/]
Metadata:
subject : Economic geography [1.000]
subject : Development studies [1.000]

```

Compare DISTIL results to manual indexing

- Encouraging initial overlap between DISTIL output and manual indexing
- Less/zero matches:
 - When no subject metadata
 - Only one or two high level subject metadata terms
 - some mismatches (as compared) when key subject elements missing, or lack of subject area coverage in DDC

DISTIL DDC indexing results - key	
	= Equal to manually indexed class
	= Narrower descendant of manually indexed class
	= Broader ancestor of manually indexed class
	= Sibling of manually indexed class
	= No match with manually indexed class

intute : sosig1028557415-27547 http://www.ifs.org.uk/	DISTIL DDC indexing - top 10 results									
	1	2	3	4	5	6	7	8	9	10
	333 - Economics of land and energy	336 - Public finance	338 - Production	331 - Labor economics	332 - Financial economics	553 - Economic geology	339.5 - Macroeconomic policy	330 - Economics	005 - Computer programming, programs, data	public property, public finance, tax, commerce (trade), industrial law
Manual DDC indexing										
331 - Labor economics										
332 - Financial economics										
336 - Public finance										
336.2 - Taxes										
339.5 - Macroeconomic policy										
338.9 - Economic development and growth										
330.9 - Economic situation and conditions										

Metadata issues encountered

- Variation in quality and quantity of metadata input affects the quality of DISTIL results output
 - Subject phrases – nested structures
 - “Arts & Humanities--History--History by Era--18th Century History”
 - “History/Policy/Law”, “Anatomy / physiology / morphology”
 - Poor subject specificity
 - “General Resources”, “People”, “Places”, “Projects”, “Images”, “Science”, “Technology”
 - Likelihood of subject terms matching DDC labels
 - Codes: “artifact1200; artifact1137; artifact804;”, “pi3731”
 - no match
 - Phrases: “Keystone Color Me Healthy”, “Connecticut Butterfly Atlas Project”
 - possibly misleading match
 - Spelling:
 - (fairly rare) mistakes: “muscoskeletal”, “policytaxation”, “intertial navigation”, “filmsUKmarketing”
 - Queries built from repository metadata – sometimes UK English, but DDC uses US English: e.g. “color”, “paleontology”, “theater”, “humor”, “aluminum”, “anemia”. Match requires either fuzzy matching (introducing noise), or supplementary query terms
 - Misleading subject combinations
 - “SPACE”, “training”, “wireless networks”, “mobile technology” - metadata for <http://www.spacestudios.org.uk/> - an arts organisation

Conclusions

- DISTIL process run on 100,000+ repository records. Results fed back into DRAMS/SOLR system. Good basis for further refinements to basic method
- Variable quality and quantity of existing subject metadata inevitably affects quality of results
- Consider evaluation methodology and what can usefully be achieved from comparison with manual. Different options ... (e.g. how many classes?)
- How will we use DISTIL results? E.g. Recall or Precision enhancing (apply threshold)?
- Probably focus on best matching DDC classes without attempting synthetic DDC classes (e.g. combining place/date into subject)

Next steps...

- Incorporating pre-processing NLP output as DISTIL input
 - Weighted terms from titles and descriptions
- Further refinements to method
- Referencing OCLC ‘Linked Data’ class URIs
- Evaluation - objective assessment of DISTIL results
- Final run + results feedback into DRAMS / SOLR

University of
South Wales
Prifysgol
De Cymru

JISC

DISTIL

Document Indexing & Semantic Tagging Interface for Libraries

Ceri Binding
Faculty of Computing, Engineering & Science
University of South Wales

<http://hypermedia.research.southwales.ac.uk/>

ceri.binding@southwales.ac.uk



