



BNCWeb

Martin Wynne

Oxford e-Research Centre, Oxford University Computing Services &
Faculty of Linguistics, Philology and Phonetics,
University of Oxford
martin.wynne@oucs.ox.ac.uk

EGI.eu Federated Cloud Task Force 'Plugfest'
Amsterdam
12th July 2012

BNCWeb



BNCWeb is an interface to the British National Corpus, a dataset of 100 million words, carefully sampled from a wide range of texts and conversations to provide a snapshot of British English in the late 20th century.

This is a key reference work in English studies, linguistics and language teaching and is widely used in a wide variety of computational linguistic applications.

BNCWeb offers powerful search and analysis functions for searching the text and exploiting the detailed textual metadata. The BNCWeb software is an open source project. The BNC is made available by Oxford University Computing Services on behalf of the BNC Consortium for educational and research purposes, and may not be redistributed by third parties.

As part of a plan to enhance the sustainability of the resource, we aim to offer the corpus under a less restrictive licence, allowing redistribution, in the future.

The Oxford instance of the BNCWeb software is built in a VM with:

- Linux (Ubuntu 10.4 LTS 64-bit server edition)
- Apache
- Mysql
- Perl

BNCweb Query result - Mozilla Firefox

ox.ac.uk https://ota.oerc.ox.ac.uk/bncweb-cgi/main.pl?theData=[word%3D"grid"%2... clarin fcs

Your query "grid" returned 1131 hits in 383 different texts (98,313,429 words [4,048 texts]; frequency: 11.5 instances per million words)

Show Page: 1 Show Sentence View Show in random order New Query Go!

No	Filename	Hits 1 to 50	Page 1 / 23
1	A0C 766	is the GridPad, a ruggedly built, pen-operated computer made by	GRID Computer Systems. In battle
2	A0C 771	to the central system, by modem and telephone if necessary.	GRID has announced a partnershi
3	A1Y 406	gallons. Fawley power station, which supplies electricity to the National	Grid, operated at reduced capaci
4	A22 157	member of the McLaren-Honda management standing by Prost's car on the	grid in Spain while five key peop
5	A65 1117	; cemetery is 1/4 mile further on right. Roadside parking.	reference NG617207. OS ma
6	A65 1214	Coneythorpe), 5 miles W of Malton. Free parking.	Grid reference SE 707712. OS ma
7	A65 1266	Fulleby, 4 miles NE of Horncastle, by petrol station.	Grid reference TF 296733 OS ma
8	A65 1337	, on A451 and A443, 11 miles NW of Worcester.	Grid reference SO 752662 OS ma
9	A65 1409	Lane. Car park is 1/4 mile up, on left.	Grid reference 229889. OS maps
10	A65 1458	Daventry, down High Street and turn left into Church Way.	Grid reference 574548. OS maps
11	A66 1143	had a carefully worked-out policy of safeguarding power supplied through the national	grid. He forbade the Electricity B
12	A6W 495	came home fourth, 'Branca' having been stranded on the	grid when the V8 would not start
13	A6W 925	aspiration is too high. Shaun Campbell surveys the back of the	grid for grand prix racing's bigge
14	A6W 995	Prix where Trevor Taylor lined it up on 18th place on the	grid and retired during the very
15	A6X 534	so much so that Mansell was eased into fourth place on the	grid by a very impressive lap fro
16	A6X 557	McLaren which would suffer that fate. Berger was seventh on the	grid, this the result of being too c
17	A6X 583	, he failed to pre-qualify the new Jordan. Second on the	grid, second in the race, but Pros
18	A6X 592	respectively. Martin Brundle put the Brabham-Yamaha into 12th place on the	grid — despite not having the be
19	A6X 616	can spend up to 90 seconds waiting for the rest of the	grid to form. '91 Silver Arrow he
20	A6X 1188	lead after losing a third of a lap through stalling on the	grid. Rudd's spirit was returning
21	A73 2828	to him just as in another cooler season heat fanned from the	grid lower down the wall. No one
22	A75 254	supply is also regulated so that 'surges' in the National	Grid — which may accompany th
23	A77 898	'13J, Serial 18 — ZAP 205, Serial 36 —	Grid 822903 ... over.' Corporal Br
24	A77 1319	'ATO request, possible mortar attack on RUC Cookstown, ICP	Grid 483768.' The room suddenly
25	A79 1152	was a six-storey flour mill of around 1850. Internally, a	grid of cast-iron cruciform-plan c
26	A7F 751	, plant room or wherever, and the hotel comes off the	grid supply of electricity — altho
27	A7F 771	CHP unit and sell the excess electricity it generates back to the	grid. In practice the price obtain
28	A7F 773	work like crazy as Murrell suggests, and use current from the	grid when demand outstrips the
29	A7F 1489	list price is £11,109 and includes a meat probe and six gastronom-sized	grid shelves. As a basic air conve
30	A8H 64	cost and strong earnings would deter predators. Government criticised on cable	grid. By Peter Large Technology
31	A8H 67	at a Whitehall seminar there was no agreement on how that national	grid of optical cables, bringing hu
32	AAB 73	that had such points of interest as nutmeg factories instead of the	grid coordinates needed in battle
33	AAL 509	approval and could within the decade be feeding electricity into the national	grid. A decision is expected befor
34	AAL 510	Act, which gives the right to sell electricity to the privatised	grid. There have been 270 bids. I
35	AC4 3162	. He sat cross-legged on the ground and stared into the black	grid on the front of the radio. 2.2

Use cases



- 1) Specialist linguistic research, using the BNC as a basic reference dataset
- 2) University classroom teaching and learning
- 3) Independent research and a reference resource for learners, citizen scholars, etc.
- 4) Federated search in the CLARIN European e-Infrastructure
- 5) Developers build additional web services on top of BNCWeb
- 6) IT providers in institutions holding licences for the BNC implement local installations of BNCWeb for local users

Use Case 1



Researchers in linguistics and other disciplines, teachers, language learners, writers and computational linguists all around the world are potential users of BNCWeb, which is a basic reference resource for the English language.

Use Case 2

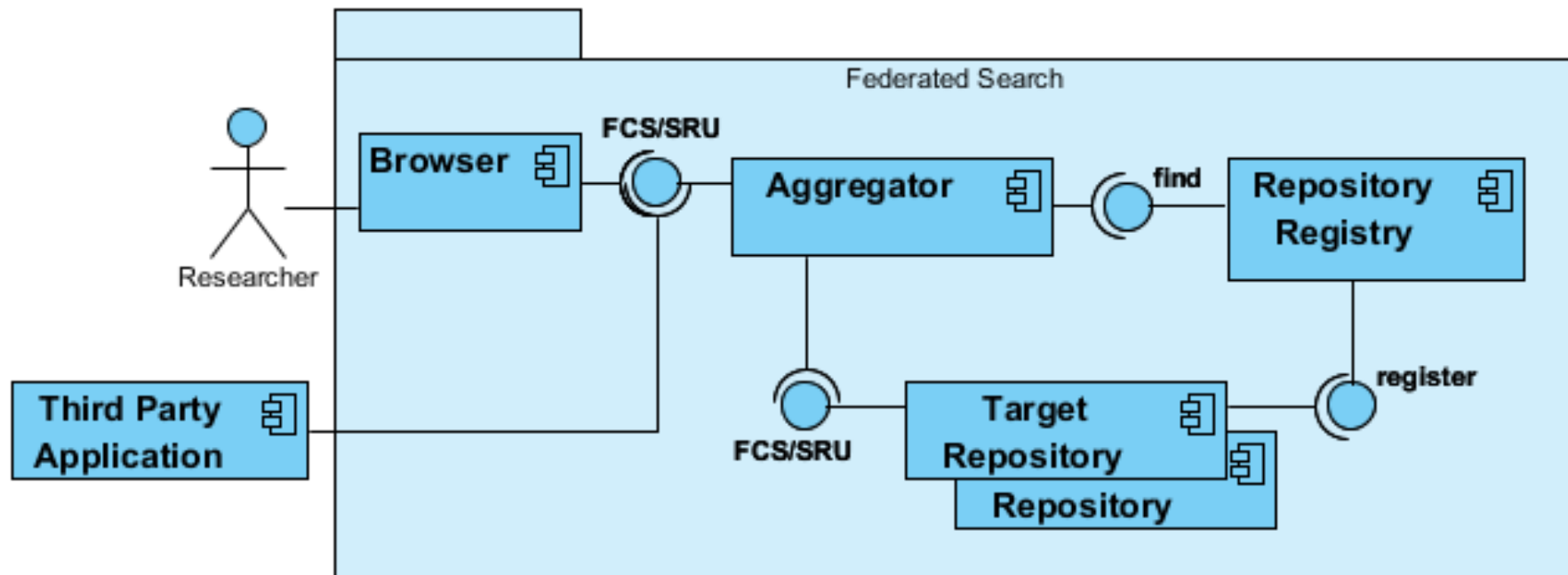


BNCWeb will be used as the main resource for teaching a Masters level course in 'Exploring English Usage' in October-November 2012, and 'Corpus Linguistics' in February-March 2013. Users will submit queries in interactive sessions with BNCWeb online. There will be usage peaks during the sessions. we want to make it available as a service for other (unscheduled) teaching sessions.

Use Case 3



Federated search in the CLARIN European e-Infrastructure: a secure and highly available BNCWeb can be used to contribute English-language resources to the ongoing project to build a Europe-wide demonstrator for federated search across archives and across access federation boundaries.



Use Case 4



Developers can build additional web services on top of BNCWeb, e.g. adding improved visualizations of the search results:

Your query "scientific" returned 5796 hits in 1165 different texts. The current solution set was found in 1165 texts. [data retrieved from cache]

Categories:
 Categories (for crosstabs only):

The following distribution was found:

Spoken or Written:

Category	No. of words	No. of hits	Dispersion (over files)	Frequency per million words
Written	87,903,571	5,665	1,120/3,140	64.45
Spoken	10,409,858	131	45/908	12.58
total	98,313,429	5,796	1,165/4,048	58.95

Derived text type:

Category	No. of words	No. of hits	Dispersion (over files)	Frequency per million words
Academic prose	15,778,028	1,787	224/497	113.26
Non-academic prose and biography	24,178,674	2,215	364/744	91.61
Unpublished written material	4,466,673	367	56/251	82.16
Other published written material	17,924,109	855	280/710	47.7
Newspapers	9,412,174	221	114/486	23.48
Other spoken material	6,175,896	126	41/755	20.4
Fiction and verse	16,143,913	220	82/452	13.63
Spoken conversation	4,233,962	5	4/153	1.18
total	98,313,429	5,796	1,165/4,048	58.95

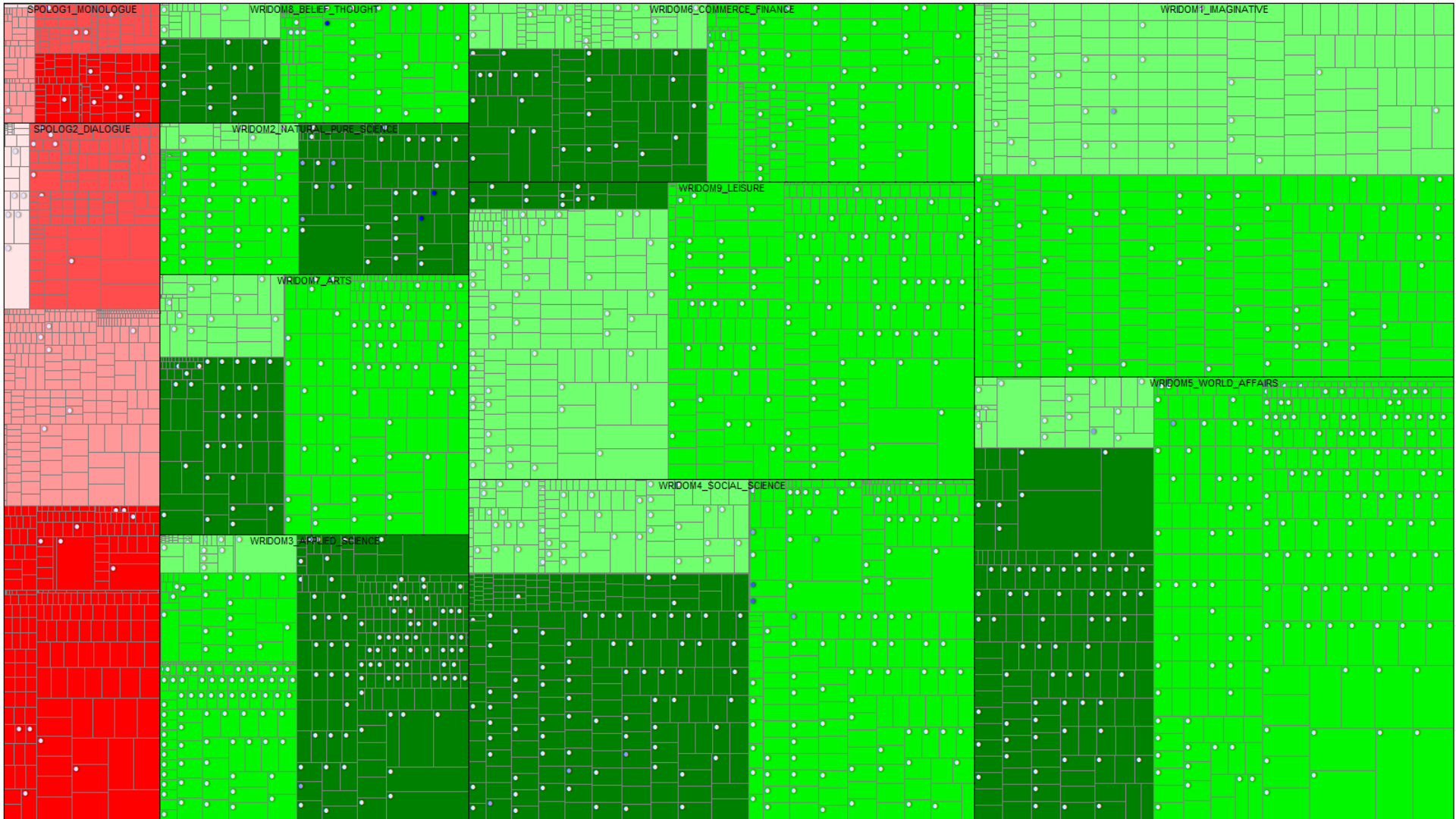
Text type:

Category	No. of words	No. of hits	Dispersion (over files)	Frequency per million words
Written miscellaneous	7,437,161	592	103/421	79.6
Written books and periodicals	79,187,792	5,049	1,003/2,684	63.76
Context-governed	6,175,896	126	41/755	20.4
Written-to-be-spoken	1,278,618	24	14/35	18.77
Demographically sampled	4,233,962	5	4/153	1.18
total	98,313,429	5,796	1,165/4,048	58.95

Text Domain:

Category	No. of words	No. of hits	Dispersion (over files)	Frequency per million words
Informative: Natural and pure sciences	3,818,803	793	70/146	207.66
Informative: Applied science	7,173,003	1,026	185/370	143.04
Informative: Belief and thought	3,037,532	397	44/146	130.7





The word "scientific" returned 5797 hits in 1165 different texts.

No. of hits: ○ 1 - 34, ● 35 - 69, ● 70 - 104, ● 105 - 138.

WRLEV0: Unknown WRLEV1: Low WRLEV2: Medium WRLEV3: High SPOREG0: Unknown SPOREG1: South SPOREG2: Midlands SPOREG3: North

Use Case 6



IT providers in institutions holding licences for the BNC implement local installations of BNCWeb for local users - e.g. <http://ota.oerc.ox.ac.uk/bncweb-cgi/BNCweb.pl/>

Requirements



Requirements:

- availability (reliable web service PLUS option for local installation)
- scalability of compute resources
- persistence (user workspace records, e.g. saved searches)
- flexible options for the access and authorization layer (basic auth / local SSO / Shibboleth)