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

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

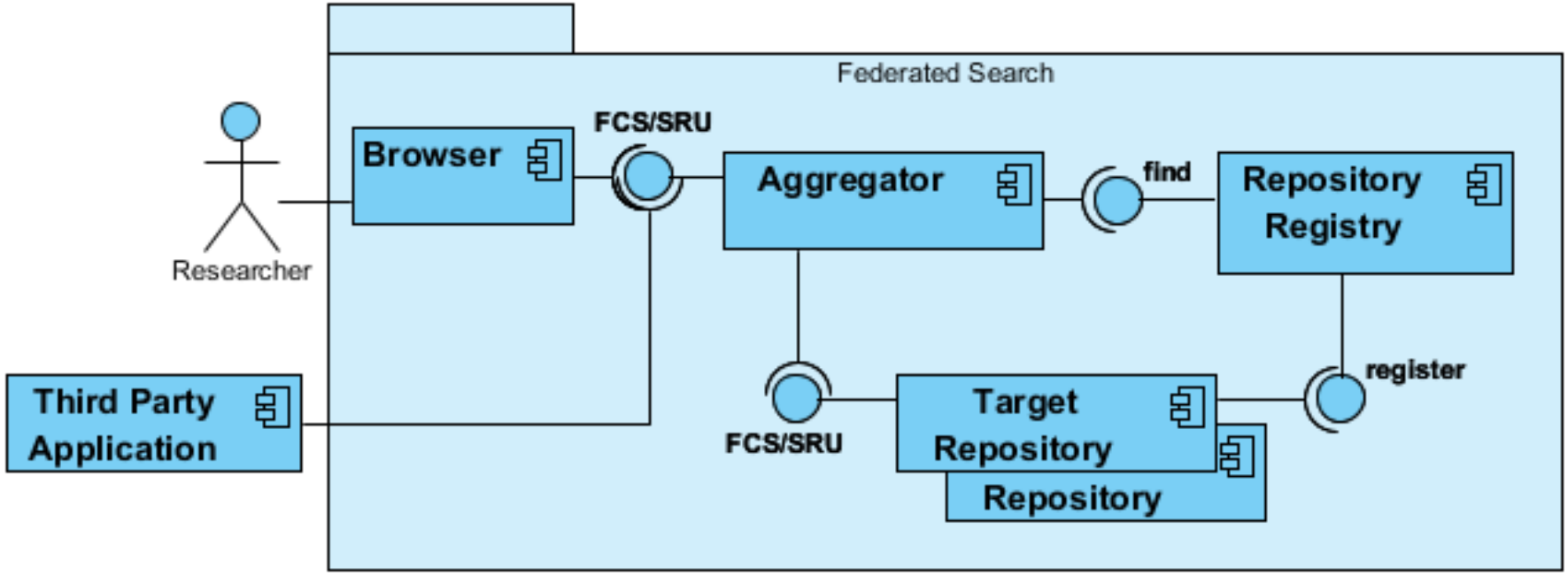
Researchers in linguistics and other disciplines, teachers, language learners, writers and computational linguists all around the world are potential users of BNCWeb, which is a basic reference resource for the English language.

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:	General information	Show distribution
Categories (for crosstabs only):	no crosstabs	New Query / Go

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:

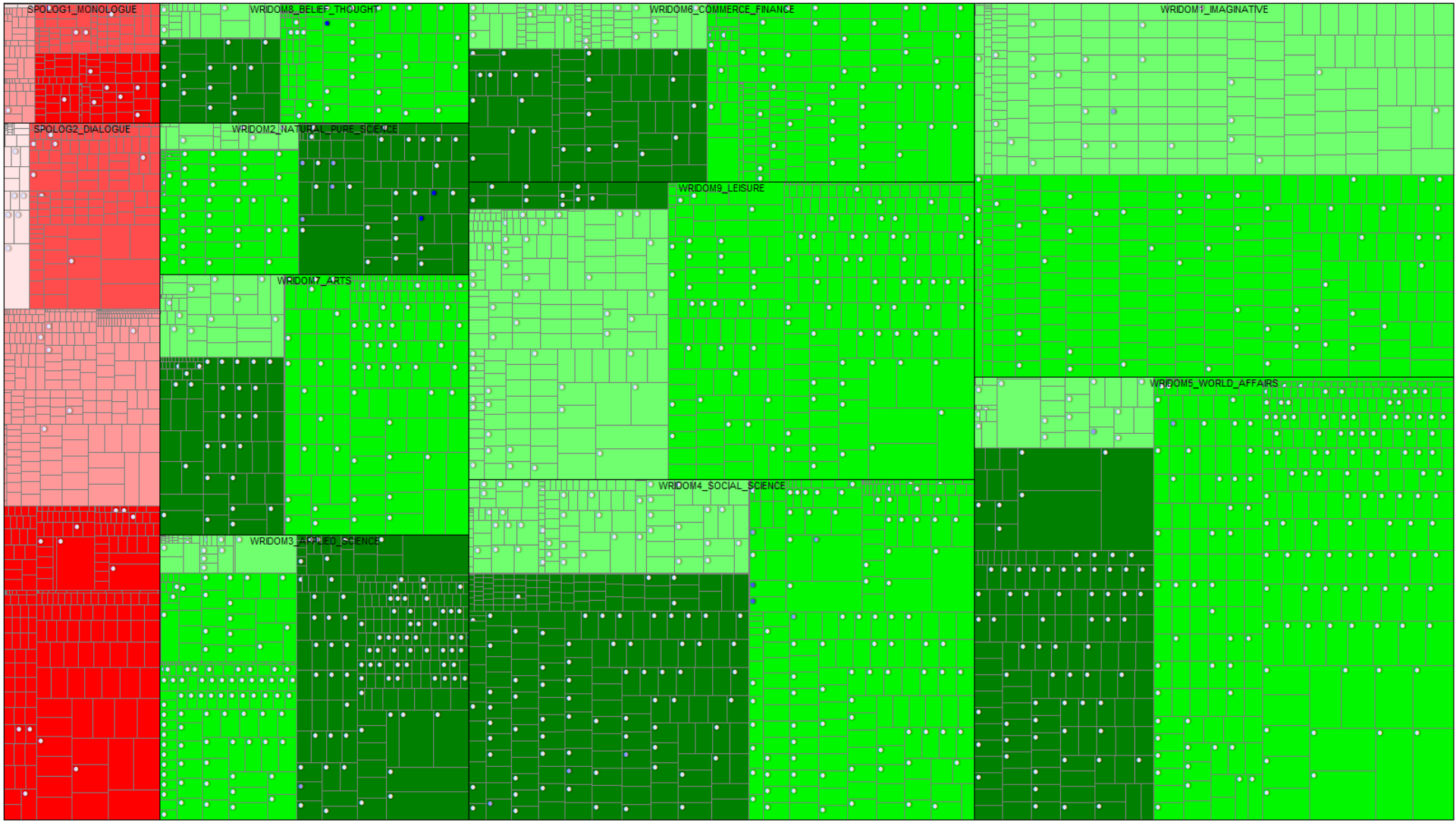
Category	No. of words	No. of hits	Dispersion (over files)	Frequency per million words
Academic prose	15,778,028	1,787	224/497	113.26
Non-academic prose and biography	24,178,674	2,215	364/744	91.61
Unpublished written material	4,466,673	367	56/251	82.16
Other published written material	17,924,109	855	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.

WRILEV0: Unknown | WRILEV1: Low | WRILEV2: Medium | WRILEV3: 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)