GATECloud.net: Cloud Infrastructure for Large-Scale, Open-Source Text Processing

Valentin Tablan Ian Roberts Hamish Cunningham
Kalina Bontcheva

University of Sheffield

28 September 2011

GATE is. . .

- Infrastructure for language engineering
- Mature (1995 -)
- Open source (LGPL, AGPL)
- Developed (mainly) at Sheffield
- GATE Family includes desktop and server software
- ≈20..30,000 downloads / year
- `http://gate.ac.uk`

Once Upon a Time. . .

The year was 2007. . .

- MATRIXWARE™ INFORMATION SERVICES
- Semantic Annotation of patents
- "We can do that!"
- "We've done large collections before (like 60,000 documents)!"

Once Upon a Time. . .

The year was 2007. . .

- MATRIXWARE™ INFORMATION SERVICES
- Semantic Annotation of patents
- "We can do that!"
- "We've done large collections before (like 60,000 documents)!"

Let's start small. . .

- ". . . with about 20 million documents"

Did you say "20 **Millions**"?!

- $\approx$ 5 sec / document
- 20 Million docs $\approx$ 3 years processing time

Fast Forward to 2009

After. . .

- two years
- a few ’00,000s €

We now had. . .

- some new software: the *GATE parallelizer*
- one IBM compute cluster (10 nodes)

. . . and were able to process the 20 million collection. . .

- in under 1 week
- using:
  - a *SUN Grid Engine* install
  - several ad-hoc shell scripts
  - and a lot of operator time

Fast Forward to 2009

After. . .

- two years
- a few '00,000s €

We now had. . .

- some new software: the *GATE parallelizer*
- one IBM compute cluster (10 nodes)

. . . and were able to process the 20 million collection. . .

- in under 1 week
- using:
  - a *SUN Grid Engine* install
  - several ad-hoc shell scripts
  - and a lot of operator time

Big data is hard!

Processing Large Collections is Useful for Scientists Too!

E.g.:

- Social Media: millions of new tweets every day
- Life Sciences: millions of Medline abstracts
- Web Research: focused web crawling can still return millions of pages
- . . .

Processing Large Collections is Useful for Scientists Too!

E.g.:

- Social Media: millions of new tweets every day
- Life Sciences: millions of Medline abstracts
- Web Research: focused web crawling can still return millions of pages
- . . .

Know any scientists that have a few ’00,000s € to spare?

2011: Introducing GATECloud.net

Design Ideas / Requirements

- Use our experience to automate large processing jobs
- Maximum automation, minimal operator intervention
- Scalability by distributing effort on a IaaS platform
- Should work with unmodified GATE pipelines

Offering

- Large processing jobs
- On-demand dedicated servers with specialised server software for text processing

GATECloud.net

Decisions & Compromises

- No AMIs (use parameterised cloud-init scripts instead)
  - (+) simpler upgrade path
  - (+) no storage costs for root partitions
  - (-) slower start-up (≈5 minutes)
- No ’elastic IPs’ (not enough addresses available); use a proxy server instead.
  - (+) unlimited names (SSL capable)
  - (-) slower throughput (proxy in the cloud could mitigate)
  - (-) single point of failure
- Sensible defaults and centralised administration instead of user configuration
  - (+) much more user-friendly
  - (-) potentially sub-optimal (but user **mis-**configuration is also a risk!)

GATECloud.net: Architectural Elements

Web App

- On-line store (item descriptions, pricing rules)
- User support (user accounts, security)
- Admin (defining new shop items, managing server swarms)
- Workflow management (orchestrating the platform)

"DNS" Proxy Server

- Secure traffic routing

The Cloud

- EC2: compute instances on demand
- SimpleDB: management of workflow tasks
- S3: software distros, job data, job results

On-demand Dedicated Servers

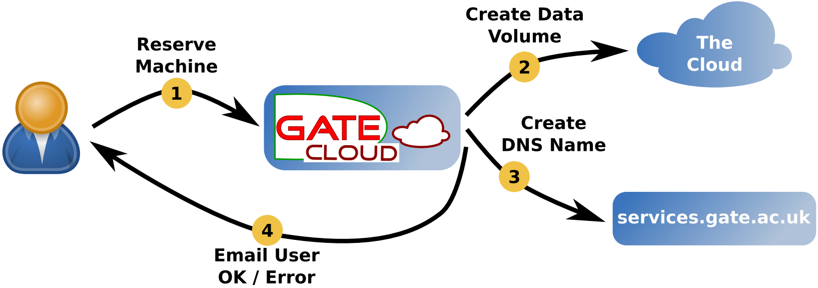
What

- GATE Teamware
- GATE Mímir

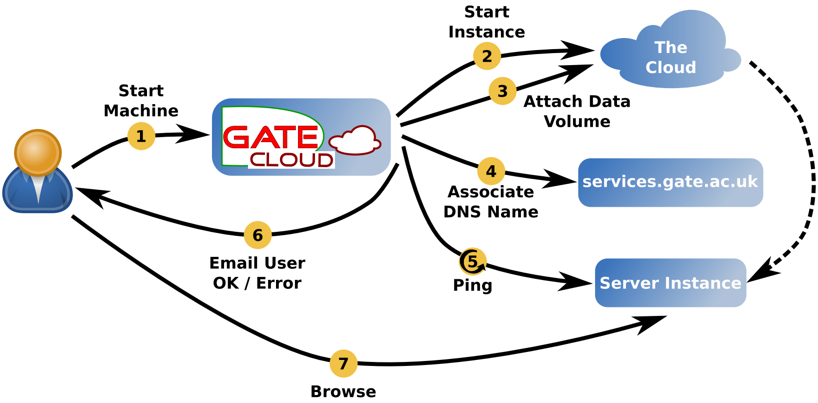
Why

- Complex build process (multiple web apps, programming languages, setup parameters)
- Complex dependencies (runtime, database setup, application server, etc.)

On-demand Servers: Making a Reservation



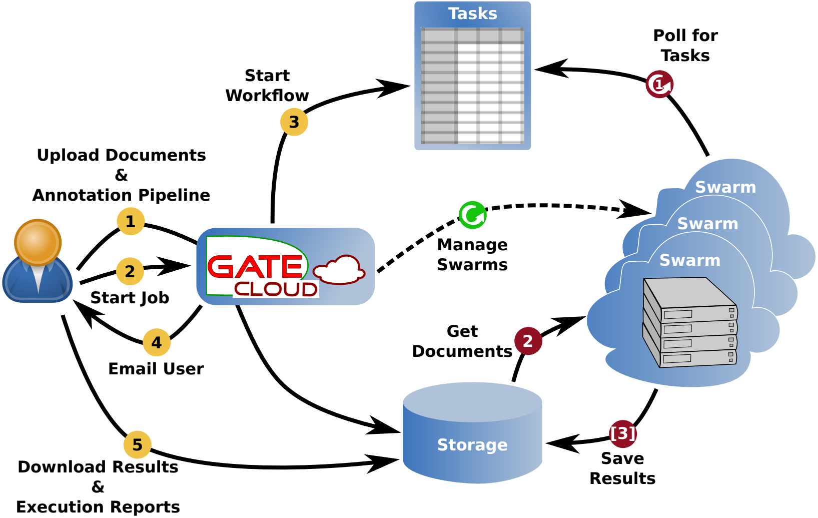
On-demand Servers: Starting a Server



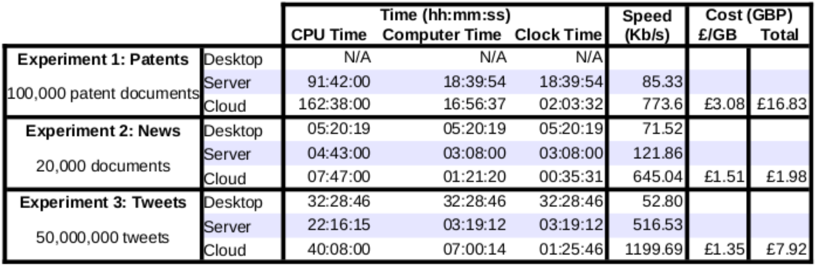
Annotation Jobs

- Process (large) document collections
- User friendly process:
  - upload documents
  - upload processing pipeline (or use pre-configured one)
  - press ’Go’
  - download results
- Maximum throughput
  - Multi-threading on each node
  - Scale up by spinning up more nodes
- Automated workflow management
- Web-based front end

Annotation Jobs: Workflow



Evaluation



		Time (hh:mm:ss)			Speed (Kb/s)	Cost (GBP)	
		CPU Time	Computer Time	Clock Time		£/GB	Total
Experiment 1: Patents 100,000 patent documents	Desktop	N/A	N/A	N/A			
	Server	91:42:00	18:39:54	18:39:54	85.33		
	Cloud	162:38:00	16:56:37	02:03:32	773.6	£3.08	£16.83
Experiment 2: News 20,000 documents	Desktop	05:20:19	05:20:19	05:20:19	71.52		
	Server	04:43:00	03:08:00	03:08:00	121.86		
	Cloud	07:47:00	01:21:20	00:35:31	645.04	£1.51	£1.98
Experiment 3: Tweets 50,000,000 tweets	Desktop	32:28:46	32:28:46	32:28:46	52.80		
	Server	22:16:15	03:19:12	03:19:12	516.53		
	Cloud	40:08:00	07:00:14	01:25:46	1199.69	£1.35	£7.92

Did it work?

20,000,000 patents:

	2007	**2009**	**2011**
Time	3 years	≈7 days	≈16 days (could be made arbitrarily shorter with a larger swarm)
Cost	N/A	00,000s € investment	≈3,500 €

Thank you!

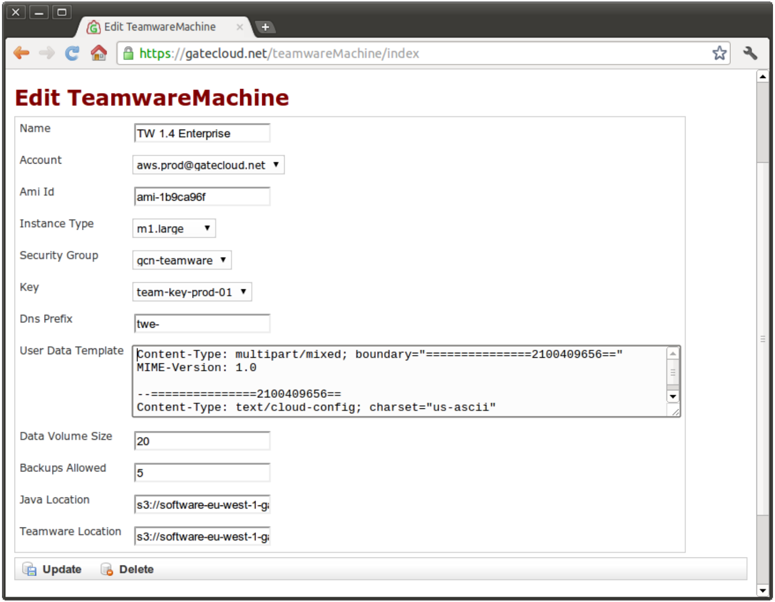
Questions?

More Information

- `http://gatecloud.net`
- `http://gate.ac.uk`

Future Work

- Investigate the use of FTL neutrinos for getting results before the job is submitted?



Edit TeamwareMachine

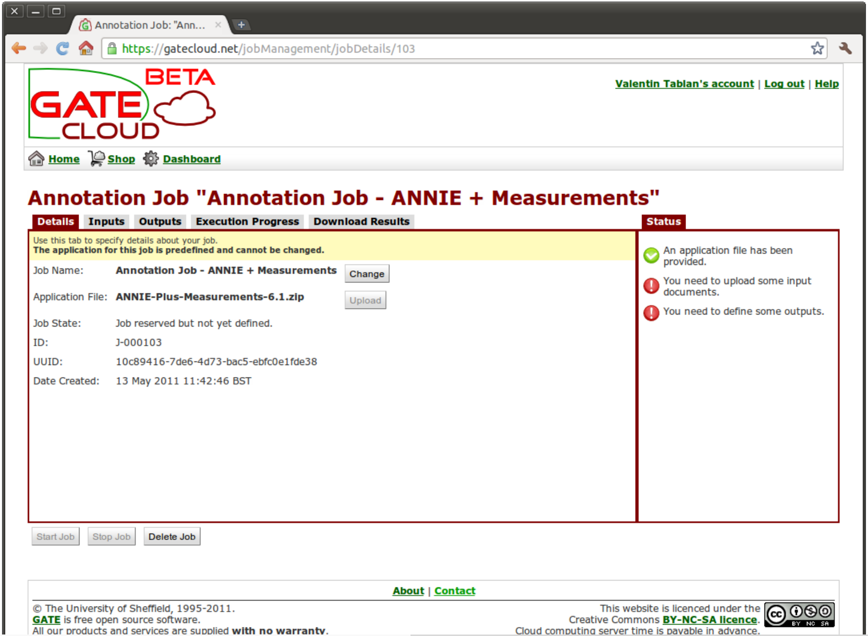
Name	TW 1.4 Enterprise
Account	aws.prod@gatecloud.net ▼
Ami Id	ami-1b9ca96f
Instance Type	m1.large ▼
Security Group	gcn-teamware ▼
Key	team-key-prod-01 ▼
Dns Prefix	twe-

User Data Template

```
Content-Type: multipart/mixed; boundary="===============2100409656=="
MIME-Version: 1.0

--===============2100409656==
Content-Type: text/cloud-config; charset="us-ascii"
```

Data Volume Size	20
Backups Allowed	5
Java Location	s3://software-eu-west-1-g
Teamware Location	s3://software-eu-west-1-g



Annotation Job: "Ann...
https://gatecloud.net/jobManagement/jobDetails/103
BETA
GATE
CLOUD
Valentin Tablan's account | Log out | Help
Home
Shop
Dashboard

Annotation Job "Annotation Job - ANNIE + Measurements"

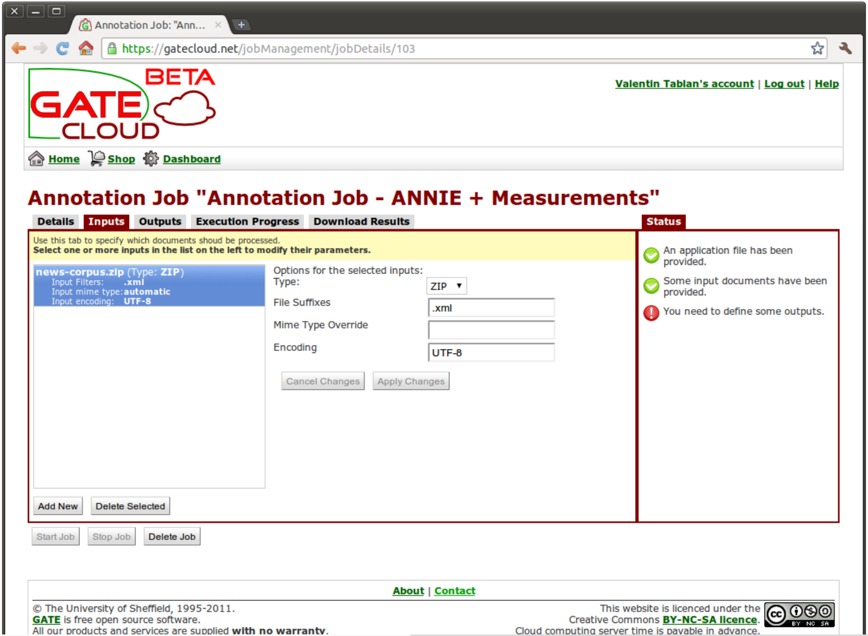
Use this tab to specify details about your job.
The application for this job is predefined and cannot be changed.

Job Name:	**Annotation Job - ANNIE + Measurements**	Change
Application File:	**ANNIE-Plus-Measurements-6.1.zip**	Upload
Job State:	Job reserved but not yet defined.	
ID:	J-000103	
UUID:	10c89416-7de6-4d73-bac5-ebfc0e1fde38	
Date Created:	13 May 2011 11:42:46 BST	

- ✓ An application file has been provided.
- ! You need to upload some input documents.
- ! You need to define some outputs.

© The University of Sheffield, 1995-2011.
GATE is free open source software.
All our products and services are supplied **with no warranty**.

This website is licenced under the Creative Commons **BY-NC-SA licence**.
Cloud computing server time is payable in advance.



Annotation Job: "Ann...
https://gatecloud.net/jobManagement/jobDetails/103
BETA
GATE
CLOUD
Valentin Tablan's account | Log out | Help
Home
Shop
Dashboard

Annotation Job "Annotation Job - ANNIE + Measurements"

Use this tab to specify which documents shoud be processed.
Select one or more inputs in the list on the left to modify their parameters.

news-corpus.zip (Type: **ZIP**)
Input Filters: **.xml**
Input mime type: **automatic**
Input encoding: **UTF-8**

Options for the selected inputs:

Type:	ZIP ▾
File Suffixes	.xml
Mime Type Override	
Encoding	UTF-8

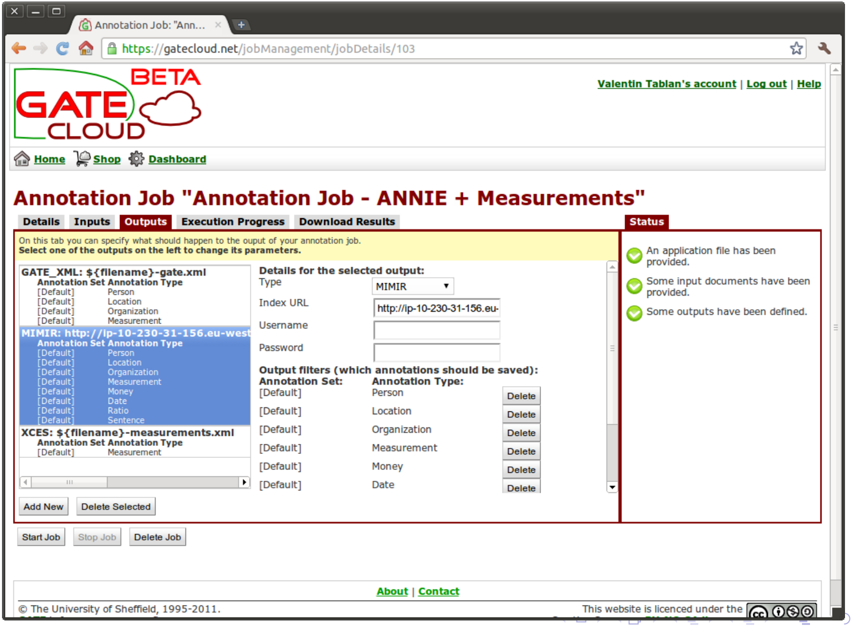
✓ An application file has been provided.

✓ Some input documents have been provided.

! You need to define some outputs.

© The University of Sheffield, 1995-2011.
GATE is free open source software.
All our products and services are supplied **with no warranty**.

This website is licenced under the
Creative Commons BY-NC-SA licence.
Cloud computing server time is payable in advance.



Annotation Job: "Ann...
https://gatecloud.net/jobManagement/jobDetails/103
BETA GATE CLOUD
Valentin Tablan's account | Log out | Help
Home Shop Dashboard

Annotation Job "Annotation Job - ANNIE + Measurements"

On this tab you can specify what should happen to the ouput of your annotation job.
Select one of the outputs on the left to change its parameters.

GATE_XML: ${filename}-gate.xml

Annotation Set	Annotation Type
[Default]	Person
[Default]	Location
[Default]	Organization
[Default]	Measurement

MIMIR: http://ip-10-230-31-156.eu-west

Annotation Set	Annotation Type
[Default]	Person
[Default]	Location
[Default]	Organization
[Default]	Measurement
[Default]	Money
[Default]	Date
[Default]	Ratio
[Default]	Sentence

XCES: ${filename}-measurements.xml

Annotation Set	Annotation Type
[Default]	Measurement

Details for the selected output:

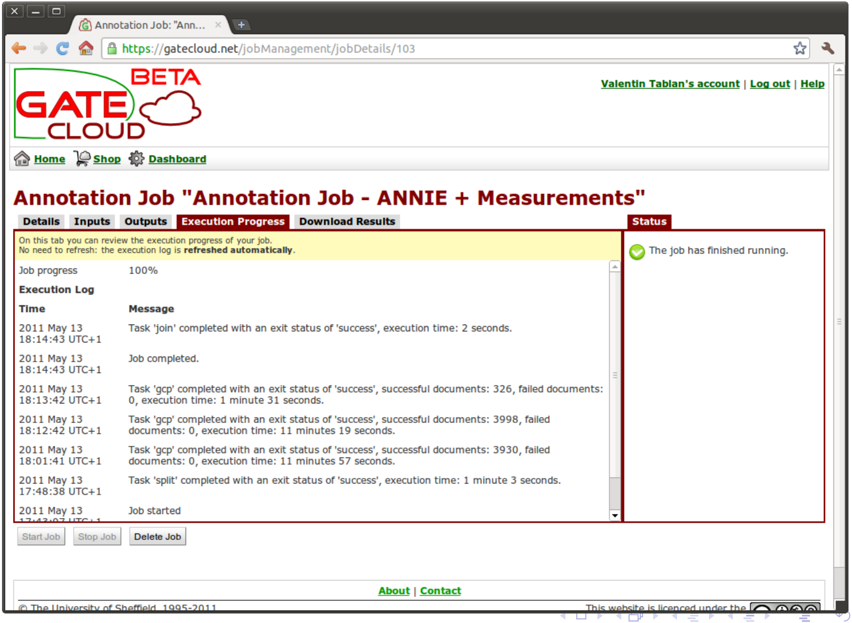
Type: MIMIR
Index URL: http://ip-10-230-31-156.eu-
Username:
Password:

Output filters (which annotations should be saved):

Annotation Set:	Annotation Type:	
[Default]	Person	Delete
[Default]	Location	Delete
[Default]	Organization	Delete
[Default]	Measurement	Delete
[Default]	Money	Delete
[Default]	Date	Delete

- ✓ An application file has been provided.
- ✓ Some input documents have been provided.
- ✓ Some outputs have been defined.

© The University of Sheffield, 1995-2011.
This website is licenced under the



Annotation Job: "Ann...
https://gatecloud.net/jobManagement/jobDetails/103
BETA GATE CLOUD
Valentin Tablan's account | Log out | Help
Home Shop Dashboard

Annotation Job "Annotation Job - ANNIE + Measurements"

On this tab you can review the execution progress of your job.
No need to refresh: the execution log is **refreshed automatically**.

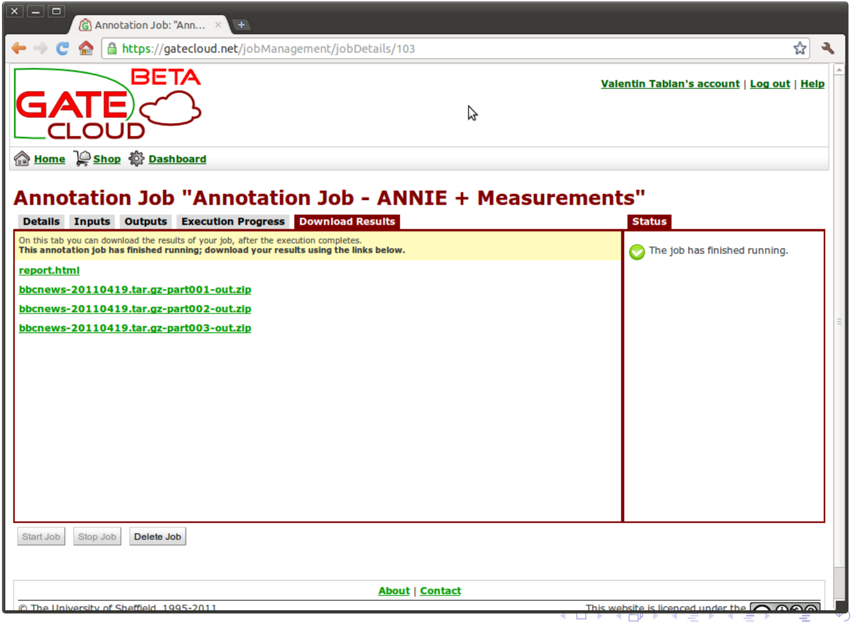
The job has finished running.

Job progress 100%

Execution Log

Time	**Message**
2011 May 13 18:14:43 UTC+1	Task 'join' completed with an exit status of 'success', execution time: 2 seconds.
2011 May 13 18:14:43 UTC+1	Job completed.
2011 May 13 18:13:42 UTC+1	Task 'gcp' completed with an exit status of 'success', successful documents: 326, failed documents: 0, execution time: 1 minute 31 seconds.
2011 May 13 18:12:42 UTC+1	Task 'gcp' completed with an exit status of 'success', successful documents: 3998, failed documents: 0, execution time: 11 minutes 19 seconds.
2011 May 13 18:01:41 UTC+1	Task 'gcp' completed with an exit status of 'success', successful documents: 3930, failed documents: 0, execution time: 11 minutes 57 seconds.
2011 May 13 17:48:38 UTC+1	Task 'split' completed with an exit status of 'success', execution time: 1 minute 3 seconds.
2011 May 13 17:43:07 UTC+1	Job started

© The University of Sheffield, 1995-2011 This website is licenced under the



Annotation Job: "Ann...
https://gatecloud.net/jobManagement/jobDetails/103
BETA GATE CLOUD
Valentin Tablan's account | Log out | Help
Home
Shop
Dashboard

Annotation Job "Annotation Job - ANNIE + Measurements"

On this tab you can download the results of your job, after the execution completes.
This annotation job has finished running; download your results using the links below.

report.html

bbcnews-20110419.tar.gz-part001-out.zip

bbcnews-20110419.tar.gz-part002-out.zip

bbcnews-20110419.tar.gz-part003-out.zip

✓ The job has finished running.

Start Job
Stop Job
Delete Job
About | Contact
© The University of Sheffield, 1995-2011
This website is licenced under the