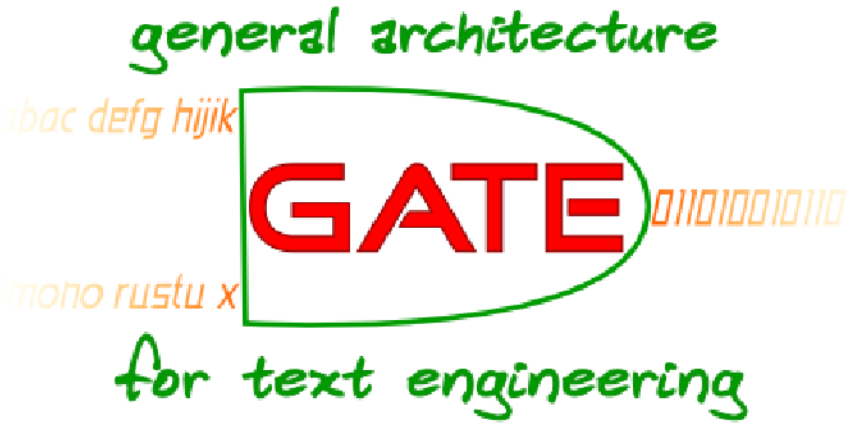
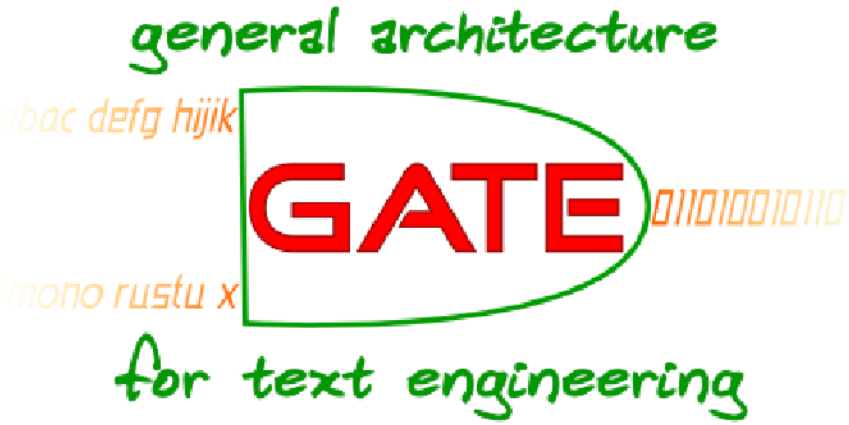


Module 4:
Taking GATE to the Cloud

GATE Teamware
GATECloud.net
GATE Mímir

© The University of Sheffield, 1995-2012
This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike Licence



Module 4:
Taking GATE to the Cloud

Part I: GATE Teamware

© The University of Sheffield, 1995-2012
This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike Licence

Outline

- Teamware
  - What is it?
  - Teamware for annotation
  - Teamware for quality assurance and curation
  - Teamware for defining workflows, running automatic services, managing annotation projects
- GateCloud.net
- Mimir

Hands-on Preparation

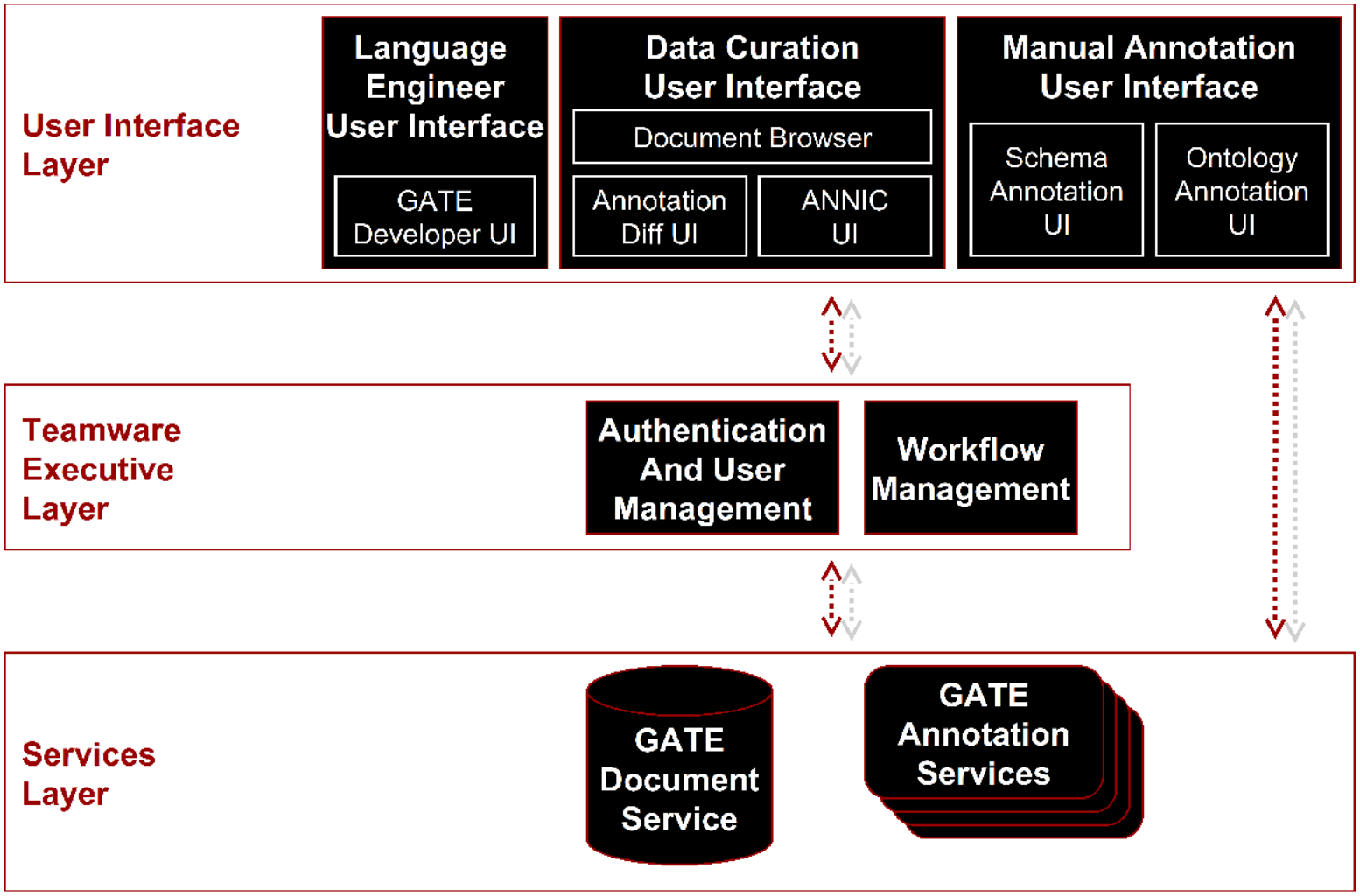
- Open the Teamware URL in your browser.
- Log on with the provided user name and password.
- Click on the link “Annotation Editor” to download and prepare the software for our first hands on. (This will get all the jars into your Java Web Start cache.)
- When it opens, leave it as is, till we need it
- Click here for more information about JWS

GATE Teamware

A Web-based Collaborative Annotation and Curation Environment

- Collaborative, distributed teams of annotators
- Structured roles for large-scale corpus annotation (annotators, managers, administrators)
- A set of generic text annotation tools for the roles
- Interactions supported in a unified environment, through workflows
- Open-source; available ready-to-use on GateCloud.net

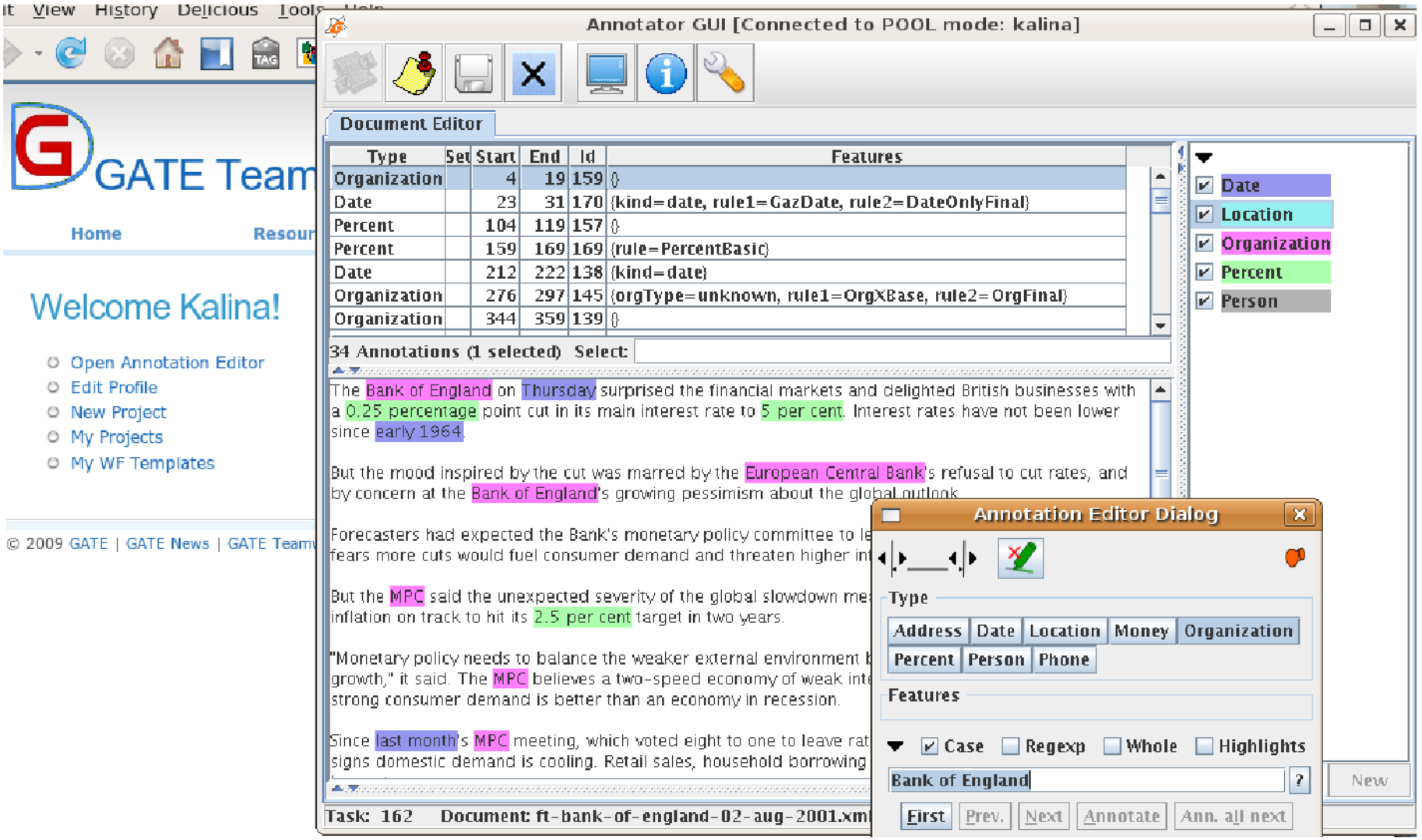
Teamware – Components



Division of Labour: A Multi-role Methodology

- **Annotators**
  - Carry out manual annotation: from scratch or (faster) by correcting automatic annotation (generated by rule-based tools, machine learning, or a combination)
- **Manager**
  - Define **annotation guidelines and schemas**
  - **Choose relevant automatic services to pre-process**
  - Toolset including **performance benchmarking**, **progress monitoring tools, small linguistic customisations**
  - Define **workflow**, **manage annotators**, liaise with language engineers and sys admins
  - **Reconcile differences** between annotators, using IAA, AnnDiff
- **Admin**
  - Setup the Teamware system, users, etc.

Teamware: Manual Annotation Tool



Manual Annotation

- Annotator: log into Teamware
- Click on “Open Annotation Editor”
- Request an annotation task (first button)
  - Annotate the assigned document
  - Save work in progress occasionally
  - Click on the “Finish task” button when done
- If you want to save work and return to this task later – “Save” button, then close the UI. Next time you request a task, you get the same document back to finish it.
- Depending on the project setup, you may be able to reject a document and then ask for another one to annotate (Reject button)

Hands-on

- Open a web browser and Teamware
- Login using your annotator user name
- Open the annotation UI
- Try requesting tasks, editing annotations, saving your work, asking for another task
  - You need to annotate names of people and organisations
- This is what Teamware looks like for manual annotation tasks

Teamware for Managers

- Defining workflows
- Running annotation projects
- Tracking progress
- Adjudication and measuring IAA

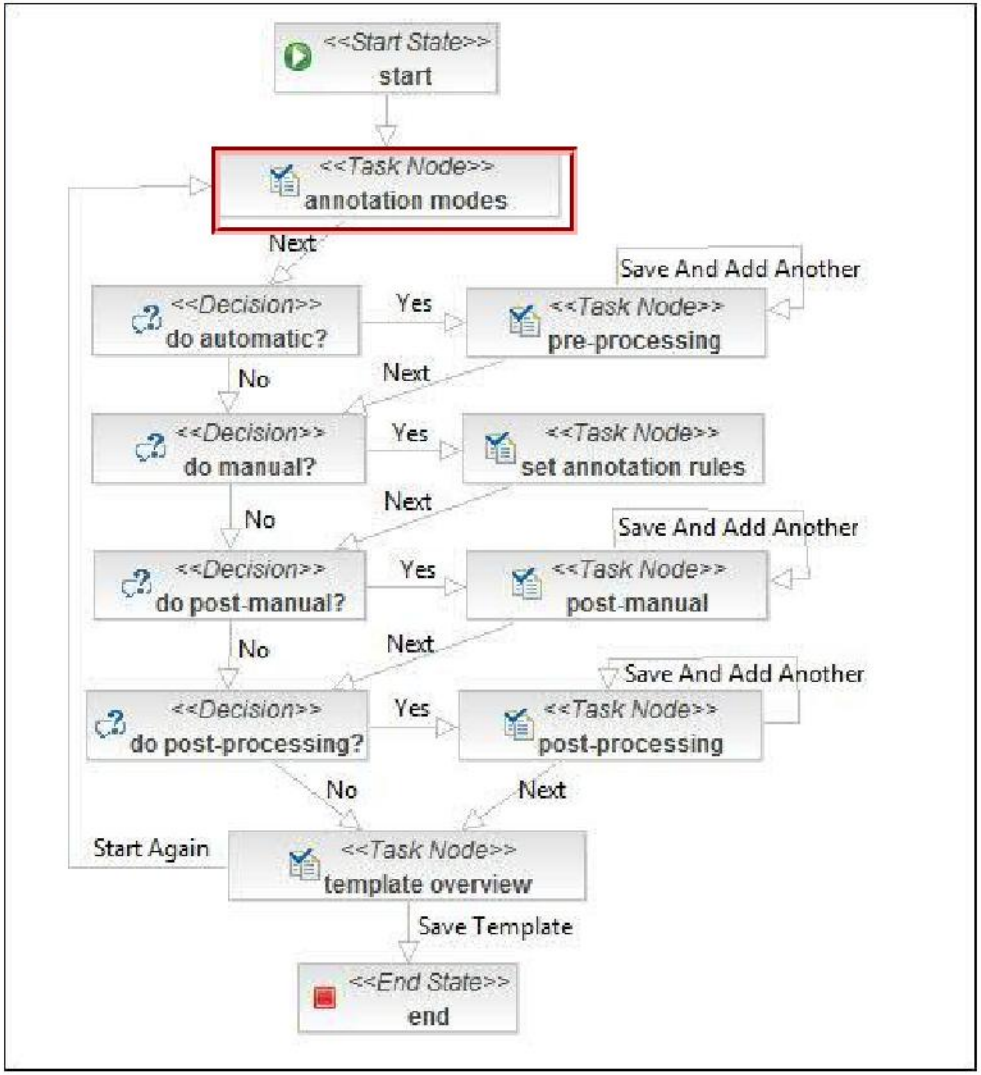
Teamware Workflows

- Whole process is controlled by a **workflow manager**
- Workflow may be simple:
  - Give the document to *n* human annotators
- or more complex
  - Pre-process the documents (Gate Developer/Gate Cloud)
  - Pass each document to 3 annotators
  - Adjudicate the differences

Workflow Templates

Automatic Annotation:	☐
Manual Annotation:	☐
Post-Manual:	☐
Review:	☐
Post-Processing:	☐

Next Help Quit

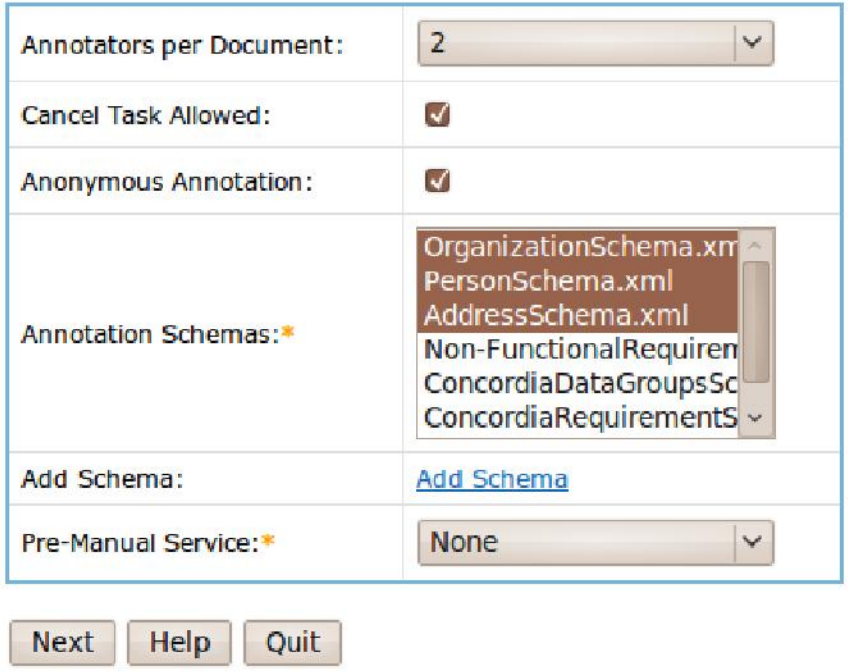


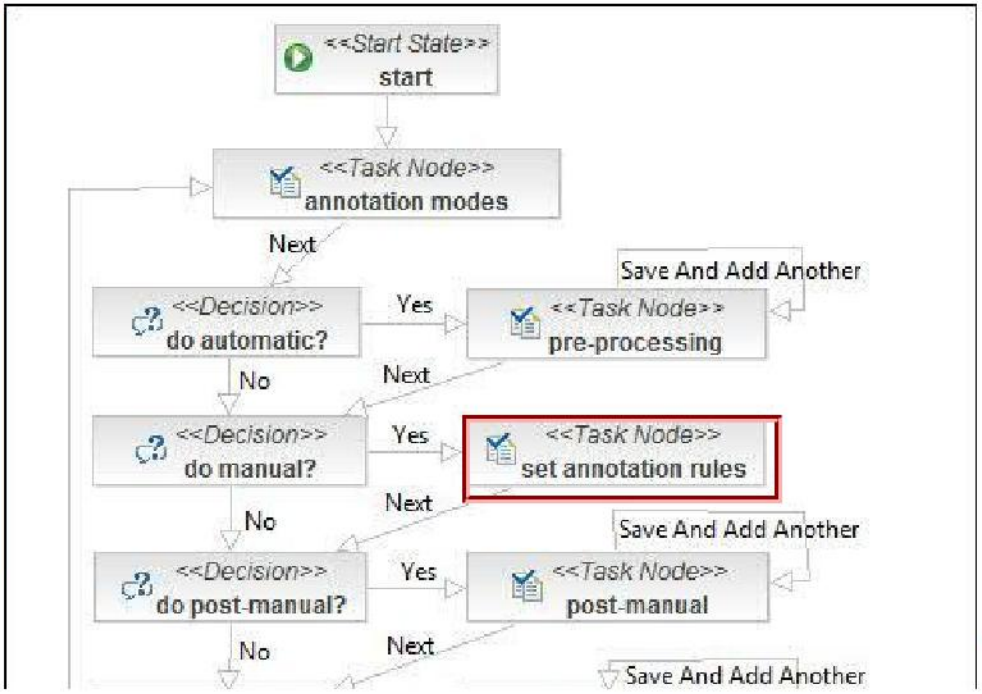
Defining new workflows

- Select Projects/WF Templates
- Opens the WF wizard
- Automatic annotation:
  - Runs one or more web services to pre-annotate
  - These need to be GATE Annotation Services (GAS), not just any service
  - The default Teamware installation includes ANNIE
- Manual annotation
- Post-manual: post-processing/merging service
- Post-processing: Finalise and export corpus

Defining Manual Annotation WF

- Select only the Manual Annotation box in the WF template
- Configure further the number of annotators per document, whether they can reject documents, whether annotation sets are anonymous (annotator1, annotator2, ...)
- Select or upload annotation schemas





Annotation Schemas

- GATE's annotation schemas can be used to constrain manual annotation for consistency
- You can use the standard ANNIE schemas or develop your own
- Schemas should be consistent with your annotation guidelines

Annotation Schemas

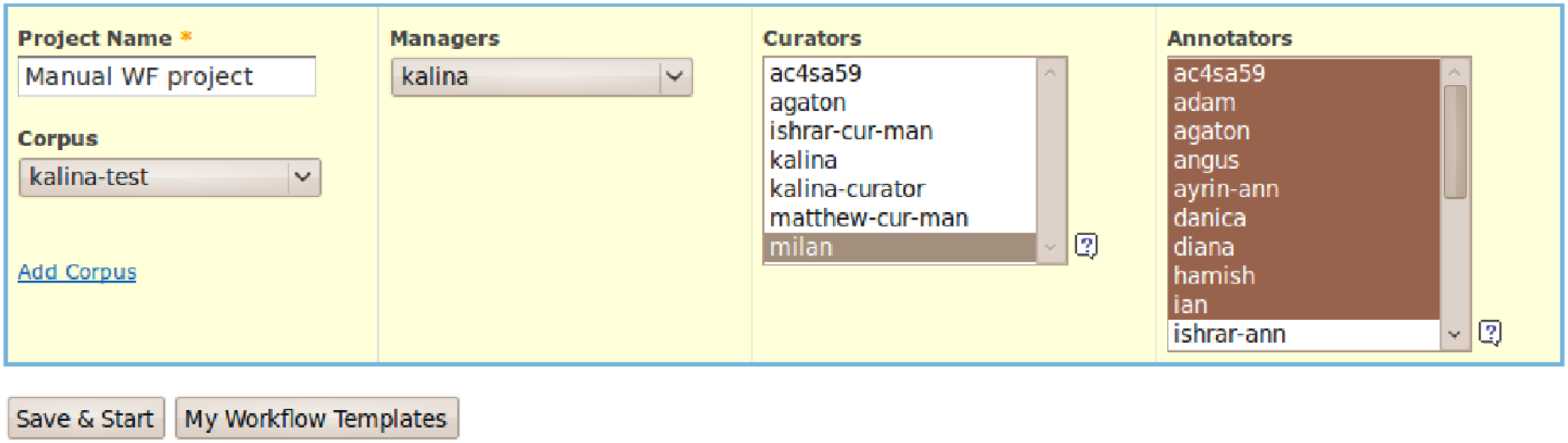
- Define types of annotations and restrict annotators to use specific feature-values
  - Person
  - Person.gender = male
  - Person.gender = female
- Uses the XML Schema language supported by W3C for these definitions

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
<element name="Person">
  <complexType>
    <attribute name="gender" use="optional">
     <simpleType>
       <restriction base="string">
         <enumeration value="male"/>
         <enumeration value="female"/>
       </restriction>
     </simpleType> ...
```

Running a Manual Annotation WF

- The WF template defined in the previous steps can be saved and the project started or revised, as necessary
- To actually run a manual annotation WF, you need to create a project, which defines:
  - Which corpus is going to be annotated
  - Who are the managers and annotators
- The corpus can be uploaded via the link provided and is a ZIP file of documents
  - can be in any format supported by GATE

Running a Manual Annotation WF



- Once project is started, annotators can login and ask for tasks

Hands-On

- Login as a manager
- Schemas:
  - Use the Person and Organization schemas provided with Teamware
- Define your own manual WF with these schemas, specifying 2 annotators per document
- Start a project with your WF template
  - Choose yourself as a manager and add your 2 annotators to the project
  - For corpus: use test-manual-workflow.zip from the resources directory. This contains 1 news document

Hands-On (Continued)

- Log out as manager and login as annotator
- Open the annotation UI
- You should now receive tasks from the projects of two other people, 2 documents in total (1 from each project – they will look identical, that's ok)
- Annotate several Person and Organization names in each task
- Remember to save and complete each task
- Make a note of any issues you had when annotating and let's discuss them

Monitoring Project Progress

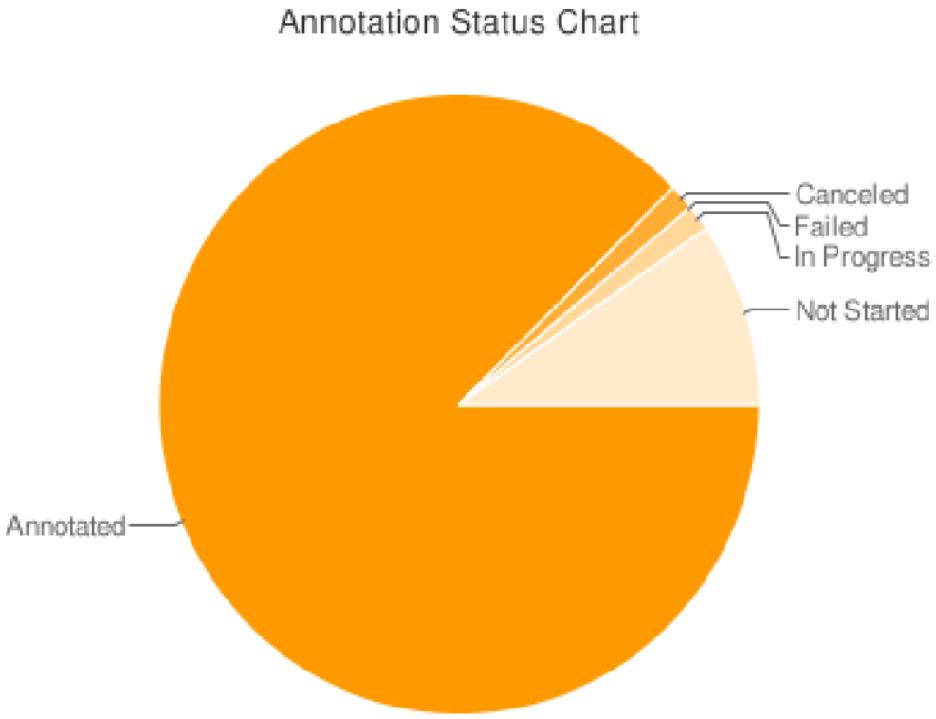
Process Monitoring: Annotation Status

Detailed View | Back to Project

Status	#
Annotated	64
Canceled	1
Failed	0
In Progress	1
Not Started	7

Average Execution Time
627.609375

Detailed View | Back to Project



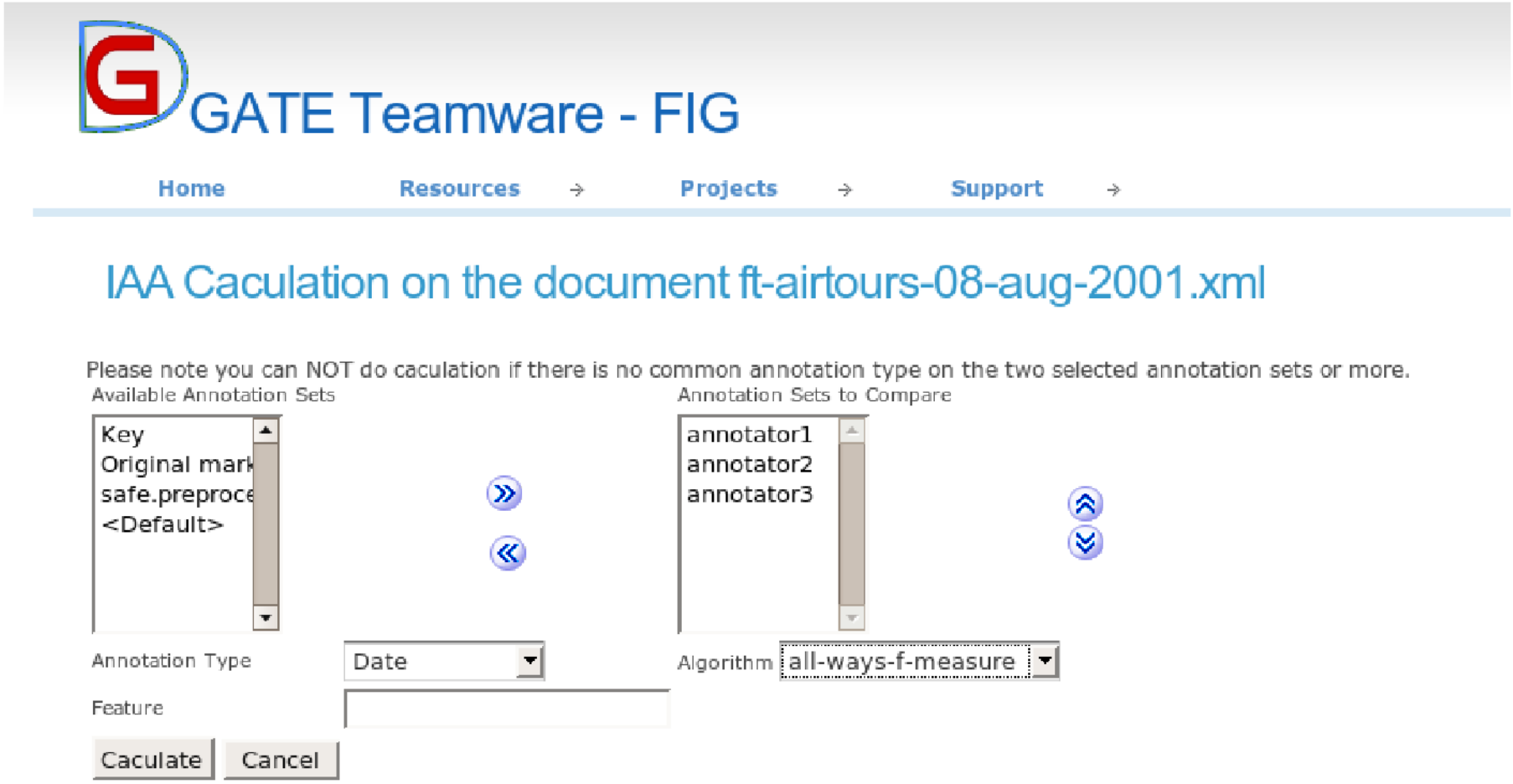
Hands-On

- Log back in as a manager
- Go to Projects/My Projects
- Select the Process Button
- You will see the manual annotation task
- Click on Monitoring to see the statistics inside
- If both your annotators have completed their tasks, you should also receive a notification email

Adjudication

- Only available to the manager role
- Identify differences between annotators using IAA
- Inspect differences in detail using AnnDiff
- Edit and reconcile differences if required
- More sophisticated adjudication UI (the Annotation Stack View) in GATE Developer

IAA: Do my annotators agree?



IAA: Results

IAA Caculation on the document ft-airtours-08-aug-2001.xml

Please note you can NOT do caculation if there is no common annotation type on the two selected annotation sets or more.

Available Annotation Sets

Key
annotator2
Original mark
annotator1
annotator3
safe.preproce
<Default>

Annotation Sets to Compare

Annotation Type: Date

Algorithm: all-ways-f-measure

Feature:

Caculate | Cancel

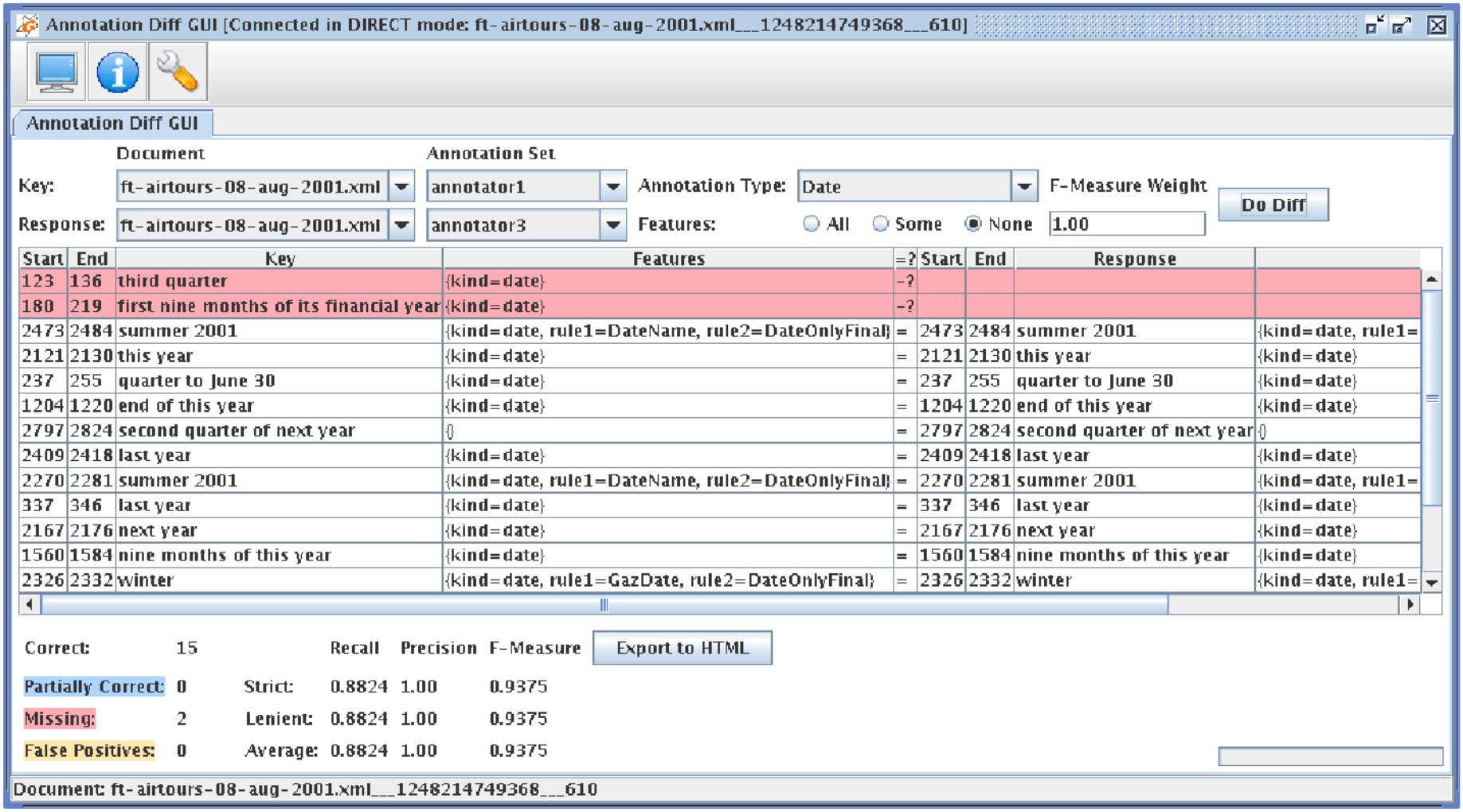
Overall F-Measure

Precision	Recall	F1	Precision Lenient	Recall Lenient	F1 Lenient	Correct	Partially Correct	Missing	Spurious
1.0	0.9411765	0.96875	1.0	0.9411765	0.96875	16.0	0.0	1.0	0.0

F-Measuse values agaist the Key annotator1

Response	SP / SR / SF	LP / LR / LF	C / P / M / S
annotator2	1.0 / 1.0 / 1.0	1.0 / 1.0 / 1.0	17.0 / 0.0 / 0.0 / 0.0
annotator3	1.0 / 0.88235295 / 0.9375	1.0 / 0.88235295 / 0.9375	15.0 / 0.0 / 2.0 / 0.0

AnnDiff: Finding the differences



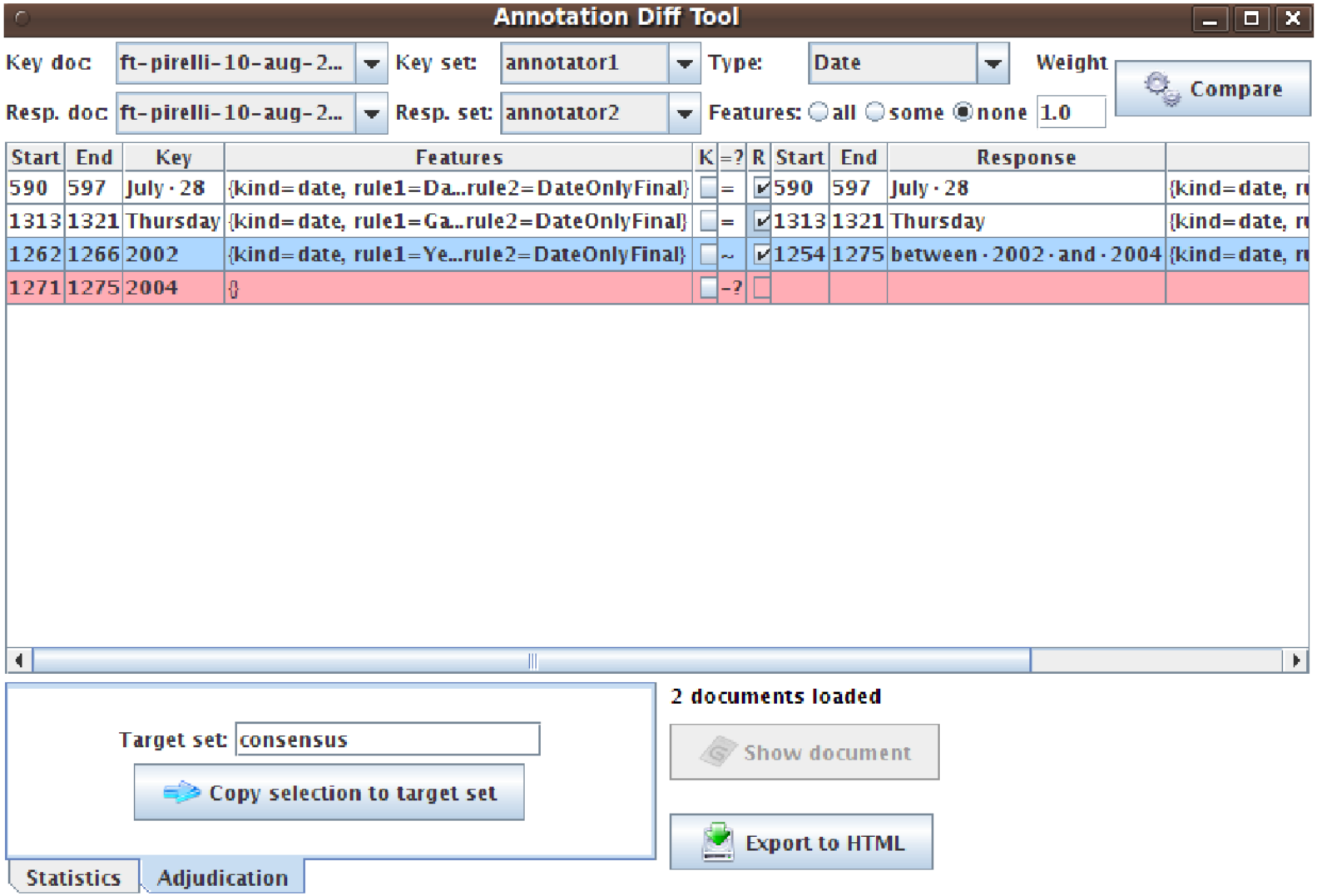
Where are these in Teamware?

- Only visible to managers
- Resources/Documents menu
- Select the corpus worked on
- Iterate through each document
- Run IAA and AnnDiff, as required
- These are clumsy, as are on a document by document basis

GATE Developer adjudication facilities

- Download the corpus as a zip file
- Corpus QA Tool
  - A corpus-level view of IAA (F-measure or Kappa)
- Extended AnnDiff to allow easy reconciliation of the differences between 2 annotators
- Annotation Stack View to reconcile difference visually, of 2 or more annotators

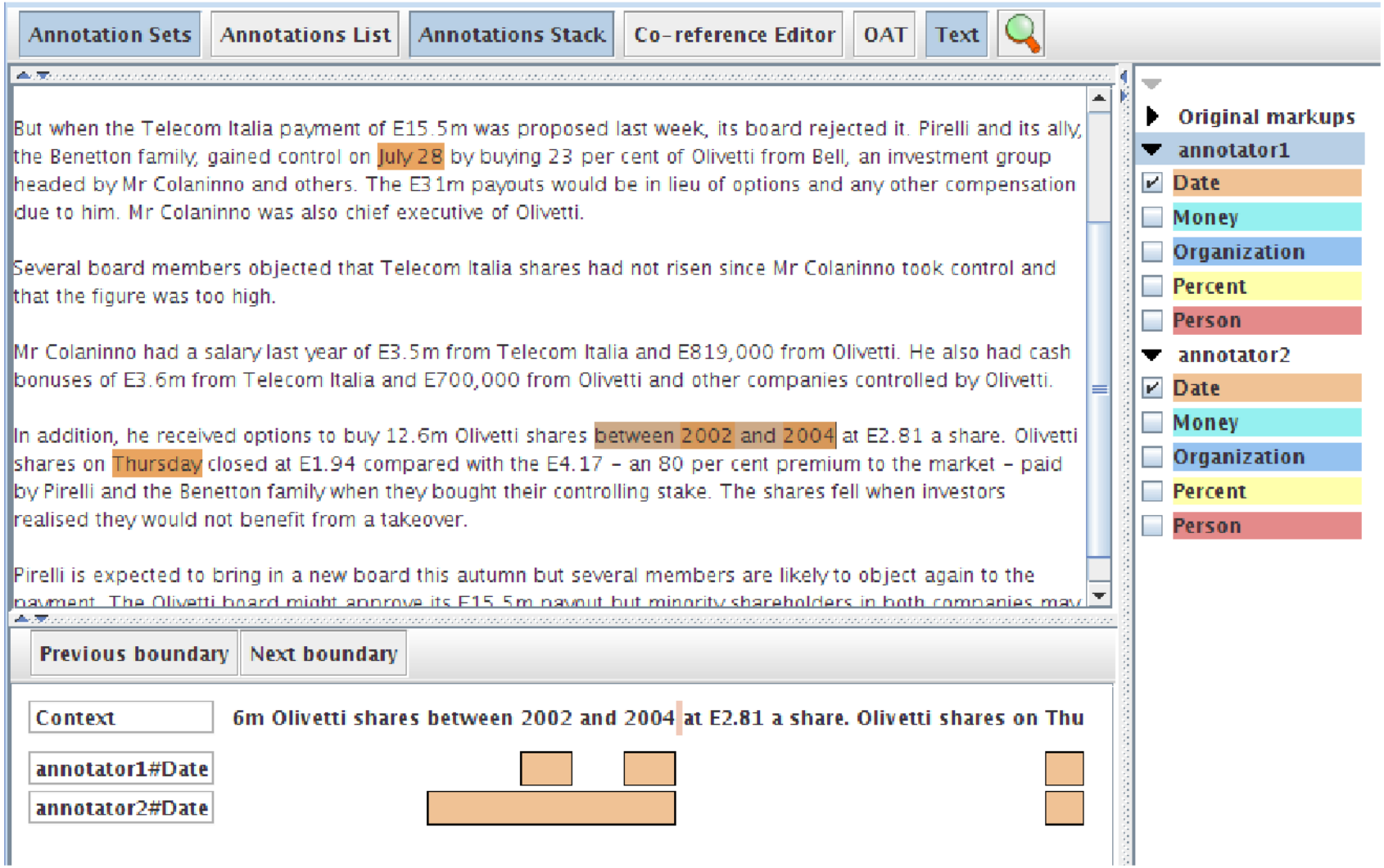
Adjudication in AnnDiff



Adjudication in AnnDiff

- Select the annotations which are correct by ticking the boxes (see previous screen shot)
- Provide the name of the target consensus set
- Click on the button to copy them into that set
- Once copied, they are removed from the list of annotations to adjudicate, so the curator can focus on the remaining ones
- Adjudication works one annotation type at a time and only for 2 annotators, whose results are stored into 2 different annotation sets

Annotation Stack View



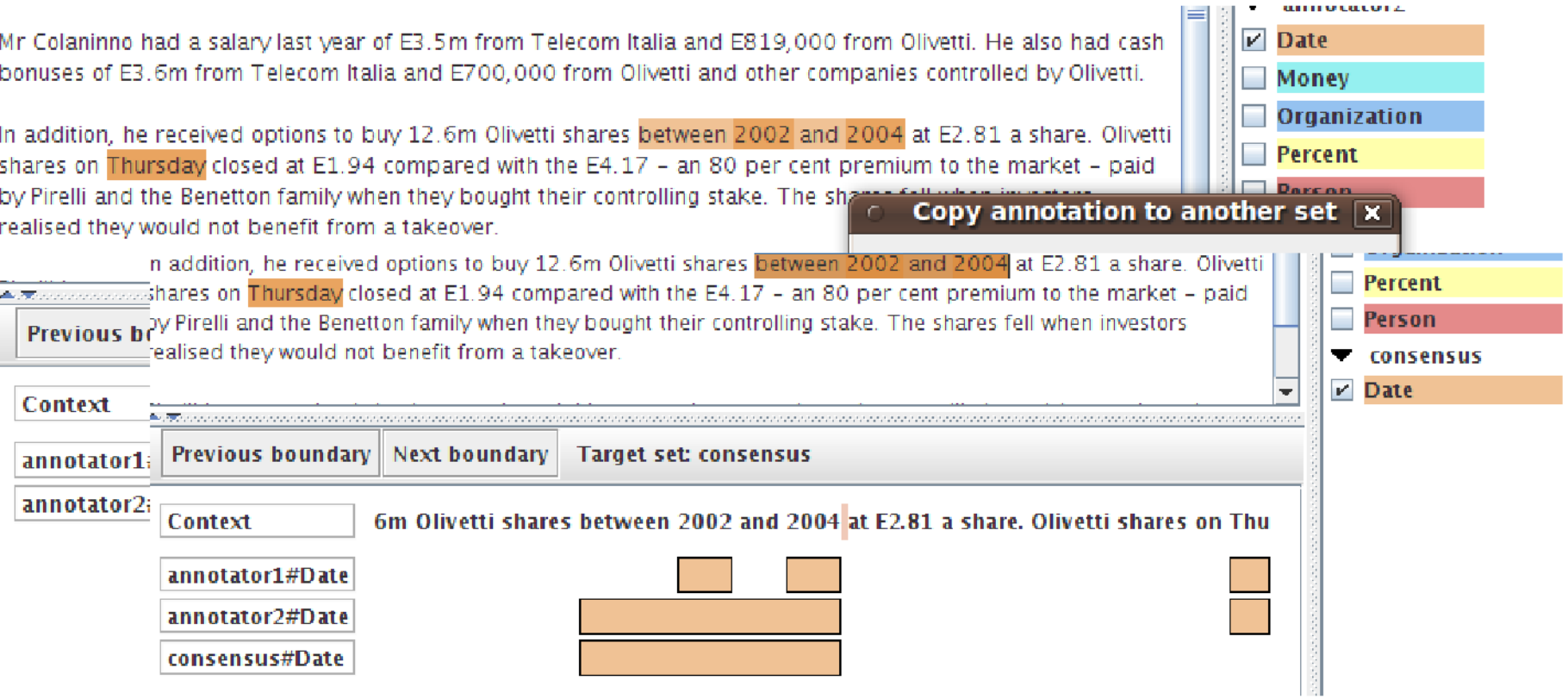
Annotation Stack View

- Open the Document Viewer
- Select the “Annotations Stack” button
- Select the annotation types you’d like to reconcile, from as many annotation sets as needed
- The Stack view window is at the bottom
- Context shows the part of the document which we are working on now

Annotation Stack View

- The Previous/Next boundary buttons position the text we're working on (the Context) on the previous/next annotation of the chosen type
- Note that such annotation may only exist in one of the sets, i.e., the other annotator might have missed it!
- Hover the mouse over the coloured annotation rectangles to see further details
- Right click would open the annotation editor
- Double click copies the annotation to the target consensus set (you choose which one it is)

Annotation Stack View



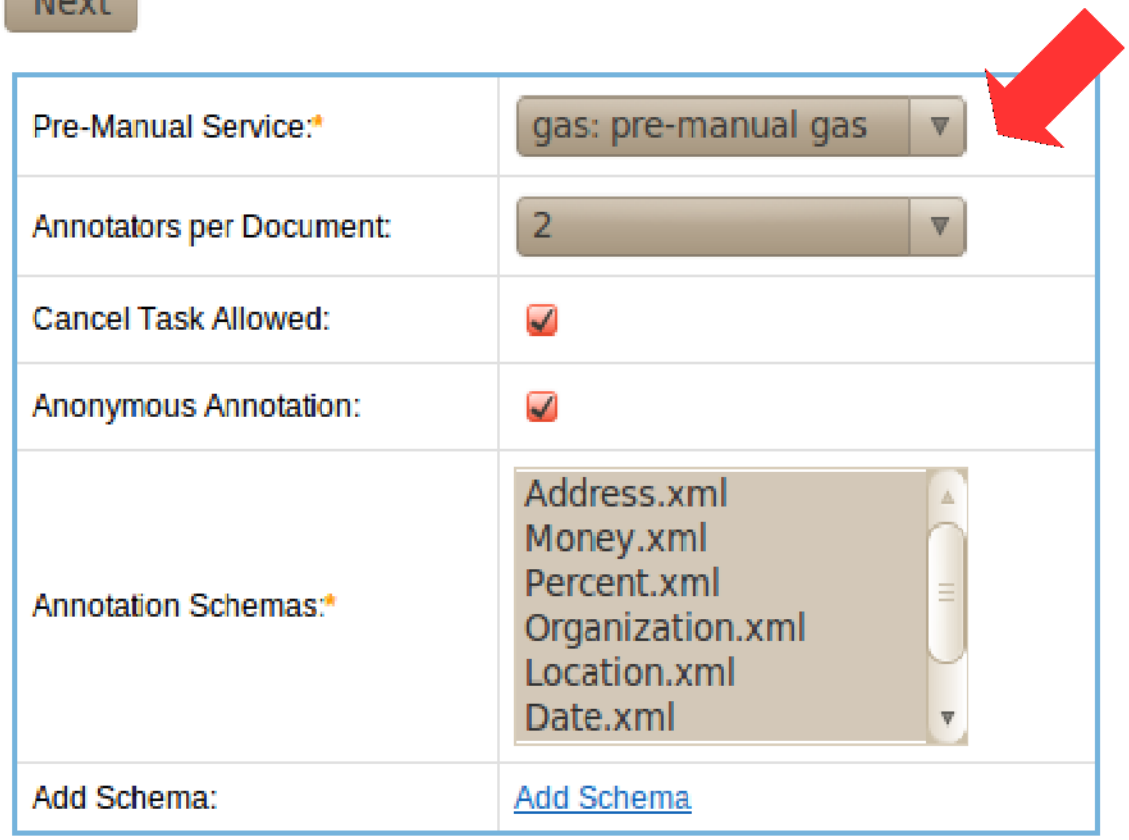
Semi-automatic Projects

- You can pre-process the corpus with pipelines in GATE Developer and then have people improve the annotations manually
- Make sure in GATE Developer that you store all automatically produced annotations in an annotation set called **safe.preprocessing**
- Then in Teamware, when you specify the manual workflow, make sure that you select to run the pre-manual service, which is called **gas:pre-manual gas**
- This service will copy all annotations from the safe.preprocessing set to the annotation set storing the results from the manual annotation, for each of the annotators. This is done at the moment when annotatorX requests this document for annotation, not earlier.

GATE

Configuration: "manual annotation"

Next



How can I use Teamware?

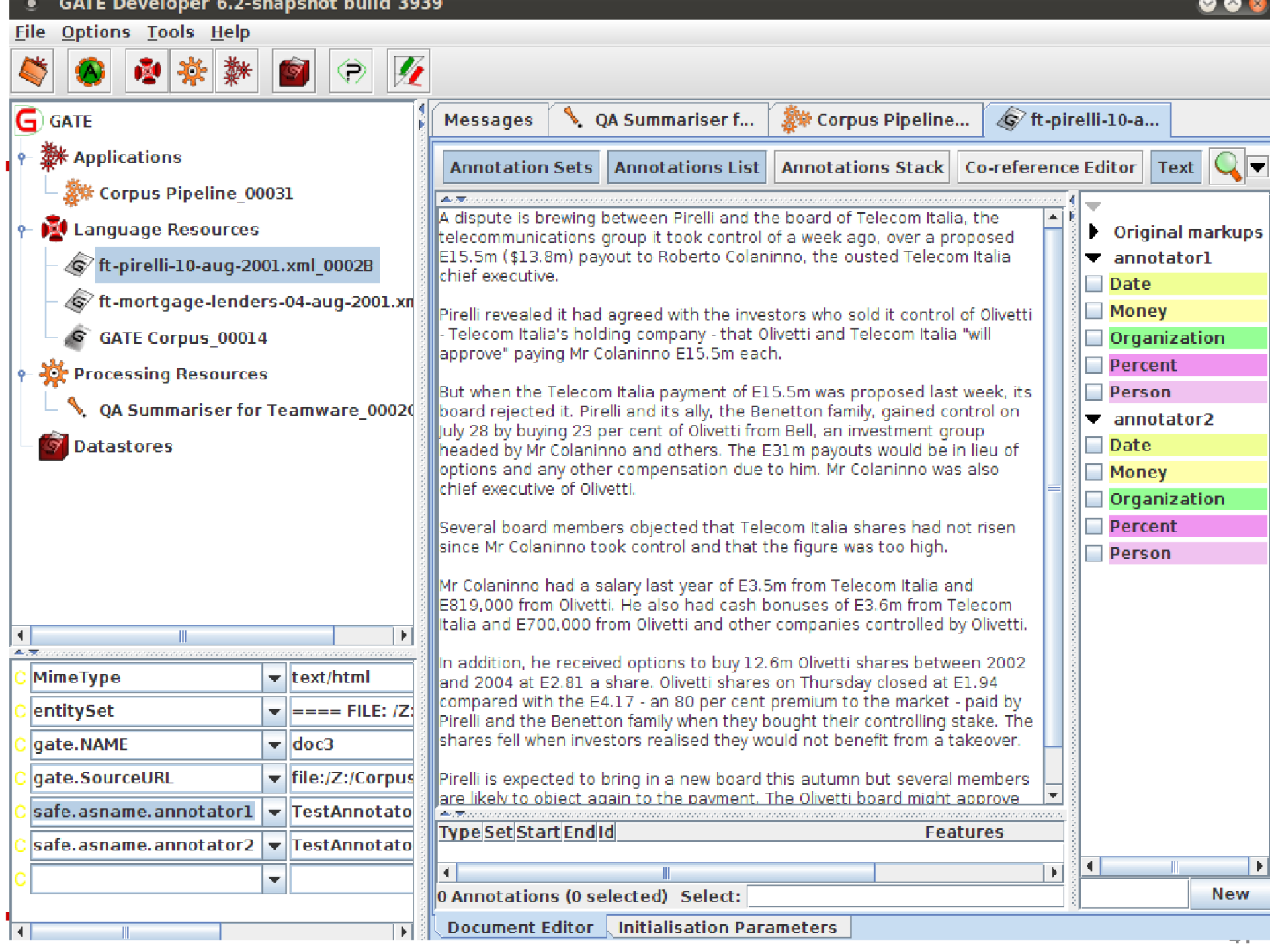
- Available from http://gatecloud.net/ for experimentation and use
- Open source from gate.ac.uk
- Contact us for a pre-installed virtual machine image, if you wish to run in-house

Advanced Teamware

- The QA Summariser
- In GATE Developer, load plugin Teamware_Tools
- It's only this tool so far, but more in the pipeline

QA Summariser

- Documents annotated in Teamware have special document features:
  - safe.asname.annotator1
  - safe.asname.annotator2
  - … (up to the number of annotators you specified per document)
- The values are the login names of the annotators who created these annotations
- Note that in one document annotator1 could be kalina and for another – mark, 3rd – danica, etc. Then kalina could be annotator2 in another document, etc.


GATE Developer 6.2-snapshot build 3939
File Options Tools Help
GATE
Applications
Corpus Pipeline_00031
Language Resources
ft-pirelli-10-aug-2001.xml_0002B
ft-mortgage-lenders-04-aug-2001.xm
GATE Corpus_00014
Processing Resources
QA Summariser for Teamware_0002(
Datastores
MimeType
text/html
entitySet
==== FILE: /Z:
gate.NAME
doc3
gate.SourceURL
file:/Z:/Corpus
safe.asname.annotator1
TestAnnotato
safe.asname.annotator2
TestAnnotato
Messages
QA Summariser f...
Corpus Pipeline...
ft-pirelli-10-a...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
A dispute is brewing between Pirelli and the board of Telecom Italia, the telecommunications group it took control of a week ago, over a proposed E15.5m ($13.8m) payout to Roberto Colaninno, the ousted Telecom Italia chief executive.
Pirelli revealed it had agreed with the investors who sold it control of Olivetti - Telecom Italia's holding company - that Olivetti and Telecom Italia "will approve" paying Mr Colaninno E15.5m each.
But when the Telecom Italia payment of E15.5m was proposed last week, its board rejected it. Pirelli and its ally, the Benetton family, gained control on July 28 by buying 23 per cent of Olivetti from Bell, an investment group headed by Mr Colaninno and others. The E31m payouts would be in lieu of options and any other compensation due to him. Mr Colaninno was also chief executive of Olivetti.
Several board members objected that Telecom Italia shares had not risen since Mr Colaninno took control and that the figure was too high.
Mr Colaninno had a salary last year of E3.5m from Telecom Italia and E819,000 from Olivetti. He also had cash bonuses of E3.6m from Telecom Italia and E700,000 from Olivetti and other companies controlled by Olivetti.
In addition, he received options to buy 12.6m Olivetti shares between 2002 and 2004 at E2.81 a share. Olivetti shares on Thursday closed at E1.94 compared with the E4.17 - an 80 per cent premium to the market - paid by Pirelli and the Benetton family when they bought their controlling stake. The shares fell when investors realised they would not benefit from a takeover.
Pirelli is expected to bring in a new board this autumn but several members are likely to object again to the payment. The Olivetti board might approve
Type Set Start End Id Features
0 Annotations (0 selected) Select:
Original markups
annotator1
Date
Money
Organization
Percent
Person
annotator2
Date
Money
Organization
Percent
Person
New
Document Editor
Initialisation Parameters

QA Summariser

- The QA Summariser lets you get performance statistics on the real annotators (e.g., kalina, mark, danica) rather than compute IAA figures for the sets named annotator1, annotator2, etc.
- The reason why we make the sets in all documents in Teamware to be called annotator1, annotator2, etc is because we want to be able to post-process the corpus easily in a unified fashion, e.g., run JAPE grammars and they all need to have the same annotation set name as input. If we have sets called by annotator login details, they will vary, e.g kalina and mark on doc1, mark and danica on doc 2

QA Summariser: Create PR

- Load the Teamware_Tools plugin
- Create a QA Summariser PR in GATE Developer (right click on the Processing Resources)
- Create a corpus pipeline and add the PR to it
- Open resources/qa-summariser-corpus.zip and unpack in a local directory
- Create a new corpus and populate it from this directory. There are 2 documents

QA Summariser: Run PR

- Provide annotation types:
  - Date, Person, Location, Organization
- No features for this example, as we are comparing only the types
- Measure: f1-average
- Provide a new output directory, where it will write the html files
- Go to that directory and open index.html in your browser

GATE

QA Summariser Output

Summary of IAA Results

AnnotationTypes: Person;Date;Location;Organization;
Features:
Measure: F1_AVERAGE

Author Names	TestAnnotator1		TestAnnotator3		Averages	
	Macro	Micro	Macro	Micro	Macro	Micro
TestAnnotator1			0.83	0.81	0.83	0.81
			document			
TestAnnotator3	0.83	0.81			0.83	0.81
	document					

Avg. IAA macro avg: 0.83
Avg. IAA micro avg: 0.81