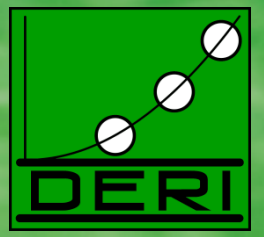


Experiences applying GATE to Semantic Web Technologies

Brian Davis,
Digital Enterprise Research Institute,
National University of Ireland, Galway
brian.davis@deri.org

DERI Institute Research Meeting, 6th Nov 08,

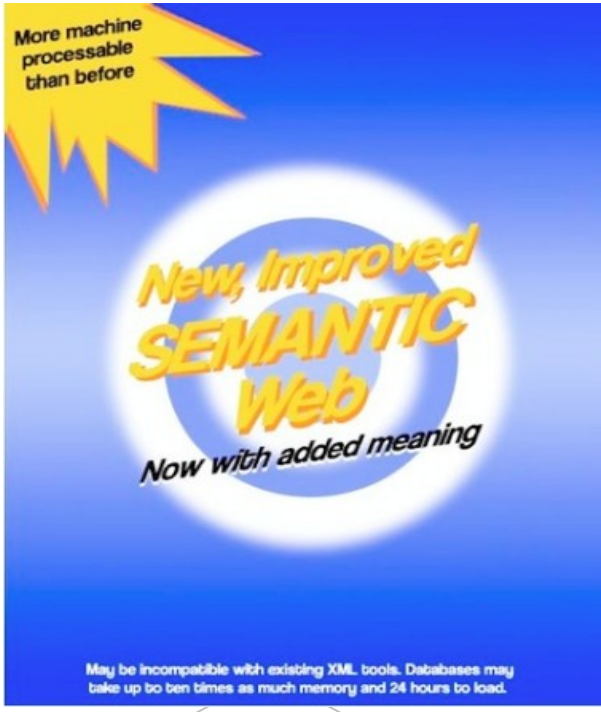
Overview

- Background
- GATE embedded to STs
  - IVEA
  - SALT
  - Semanta
- GATE as an Interface to STs
  - CLOnE, ROA
- Conclusion and Future Work

Background I: Semantic Web

Combination of standard data models and explicit semantics supports:

- information exchange and interoperability
- data integration
- improved search and retrieval
- reasoning and inference



Background I: Semantic Web

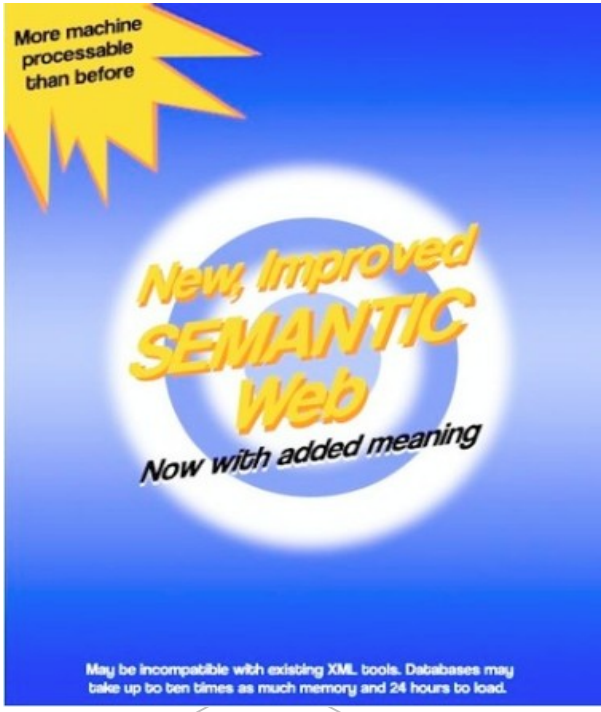
- conventional web is intended for human consumption

- content consists largely of natural language text, images, video, etc.

- Semantic Web seeks to make data more a menable to automated forms of information processing

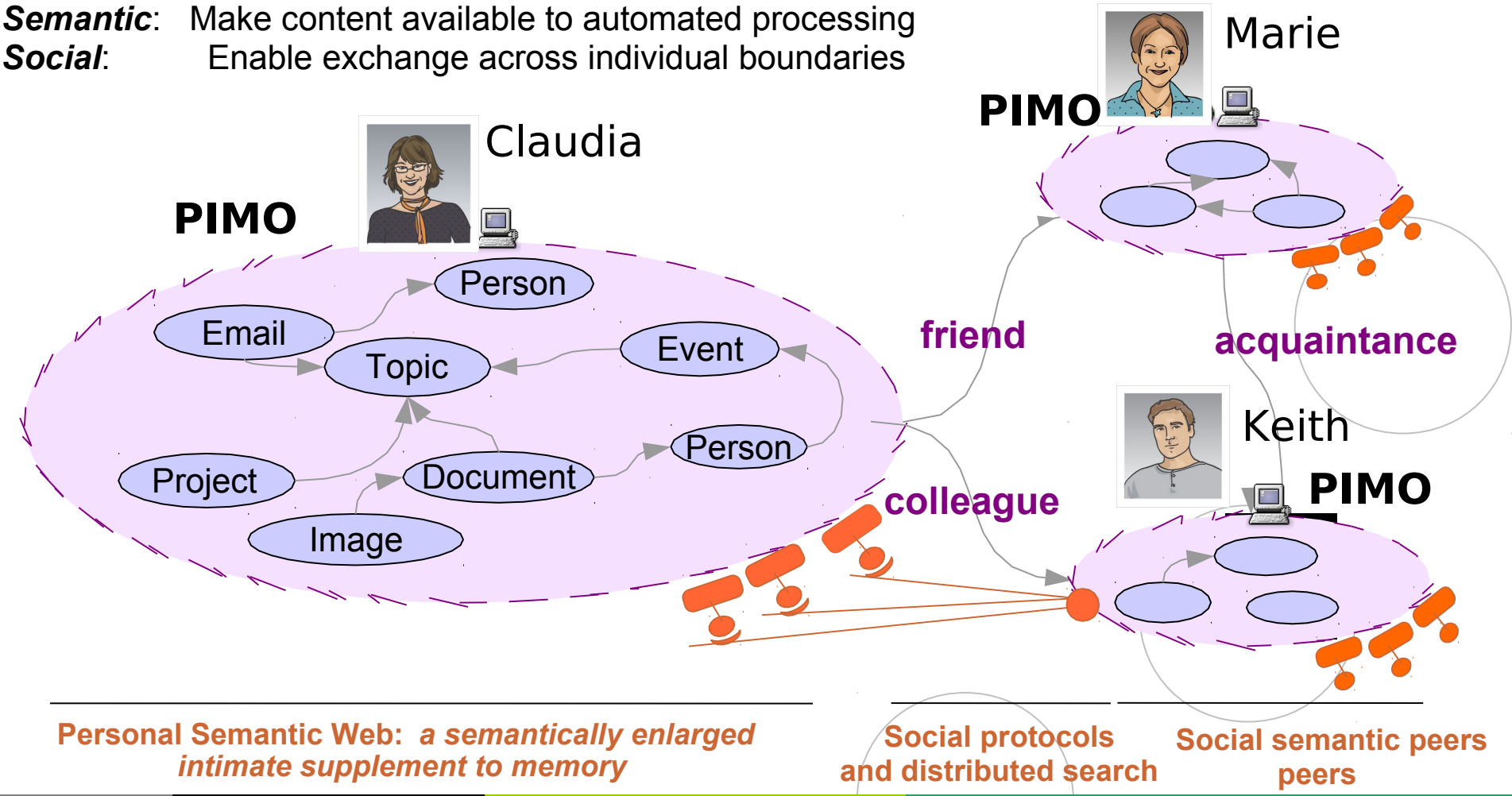
- standard data model + explicit semantics
Resource Description Framework (RDF)
core data model + some semantics

- Web Ontology Language (OWL)
more advanced semantics
OWL typically used to create ontologies that describe the conceptual structure of a specific domain of interest



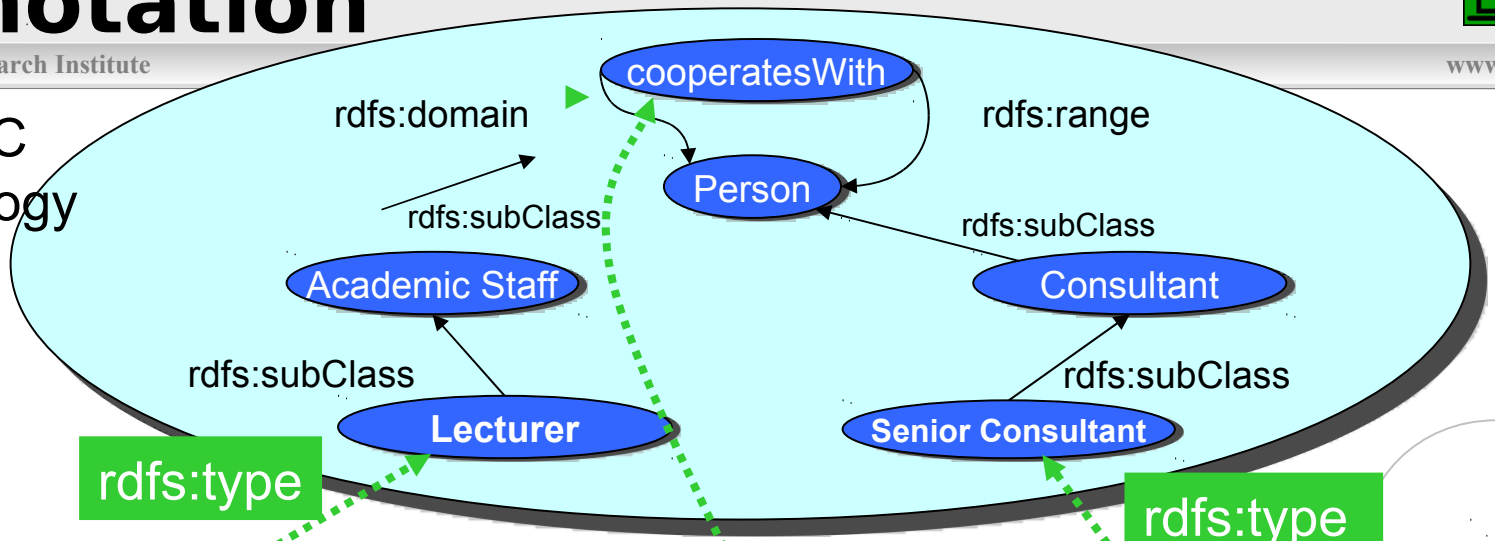
Background II: The Semantic Desktop

Desktop: Help individuals in managing information on the Web/their PC
Semantic: Make content available to automated processing
Social: Enable exchange across individual boundaries



Background III: Semantic Annotation

SWRC Ontology



rdfs:type

rdfs:type

Anno-tation

```
:sha a swrc:Lecturer;
     swrc:name "Siegfried Handschuh".

:sha swrc:cooperatesWith
  http://www.dfki.de/~declerck#tde.
```

```
:tde a swrc:SeniorConsultant;
     swrc:name "Thierry Declerck".
```

swrc:cooperatesWith

Web Page

Siegfried Handsch

He is presenting together with Therry Declerck a tutorial at ESWC 06

Links have explicit meaning!

Thierry Declerck

Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany

Building 43.8 Room +1.12

Tel: +49 681 302 5358 (office)
Fax: +49 681 302 5338
Tel: +49 681 9656634 (home)

email: declerck@dfki.de

URL

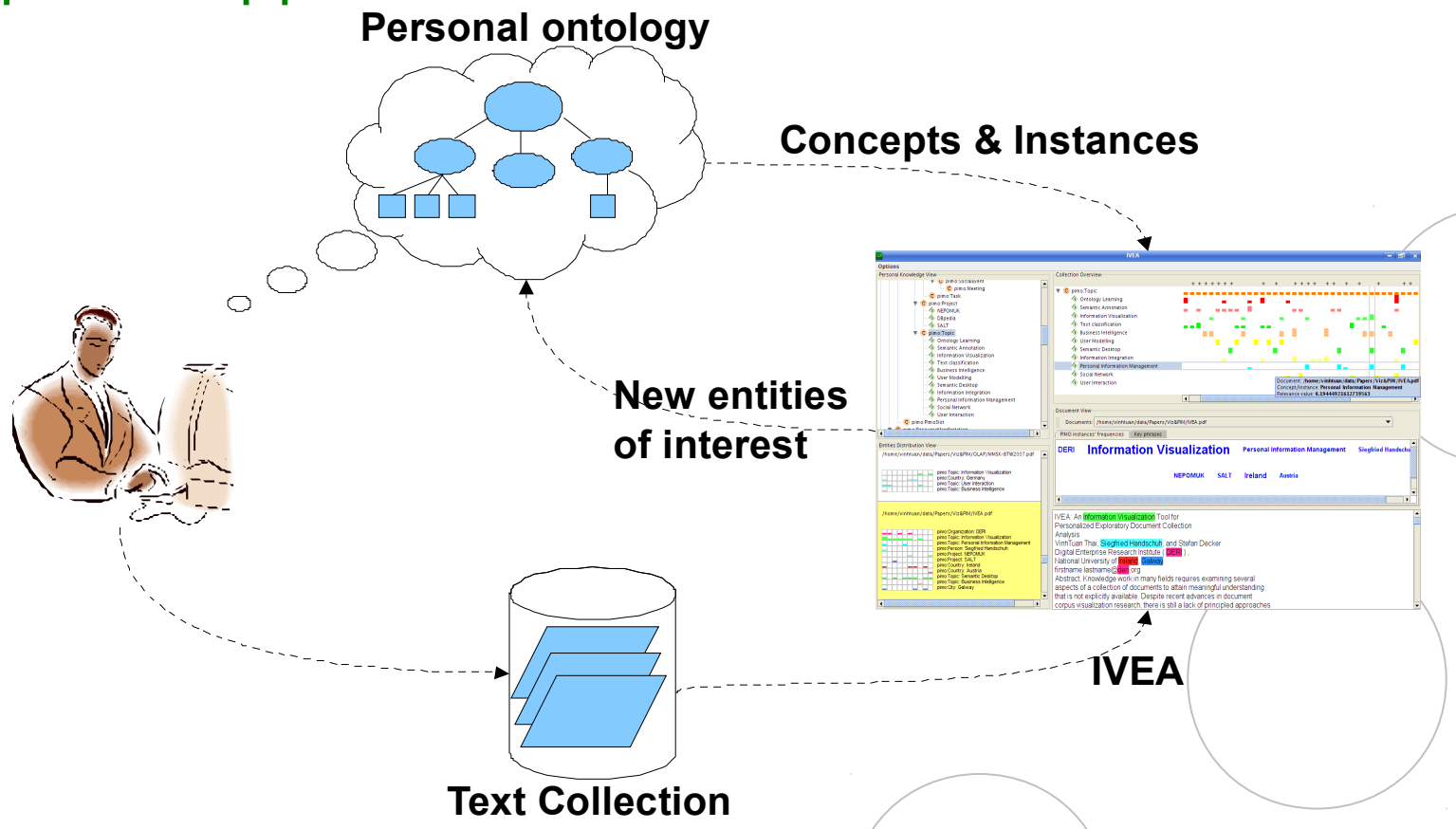
http://www.siegfried-handschuh.net

http://www.dfki.de/~declerck/

Embedding GATE in Semantic Web Applications I: IVEA

science foundation ireland
fondúireacht eolaíochta éireann
NUI Galway
OÉ Gaillimh
Enabling networked knowledge.

Proposed approach

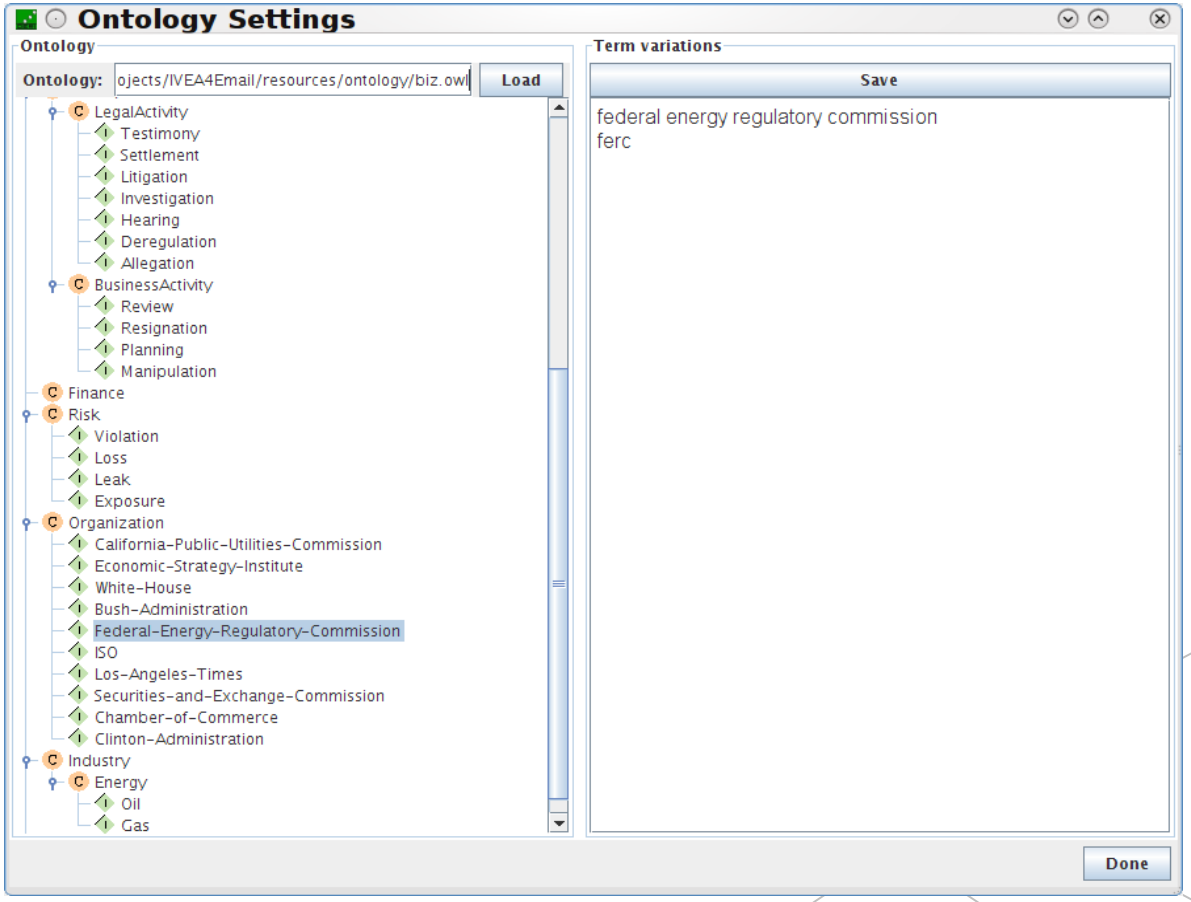


IVEA: Introduction II

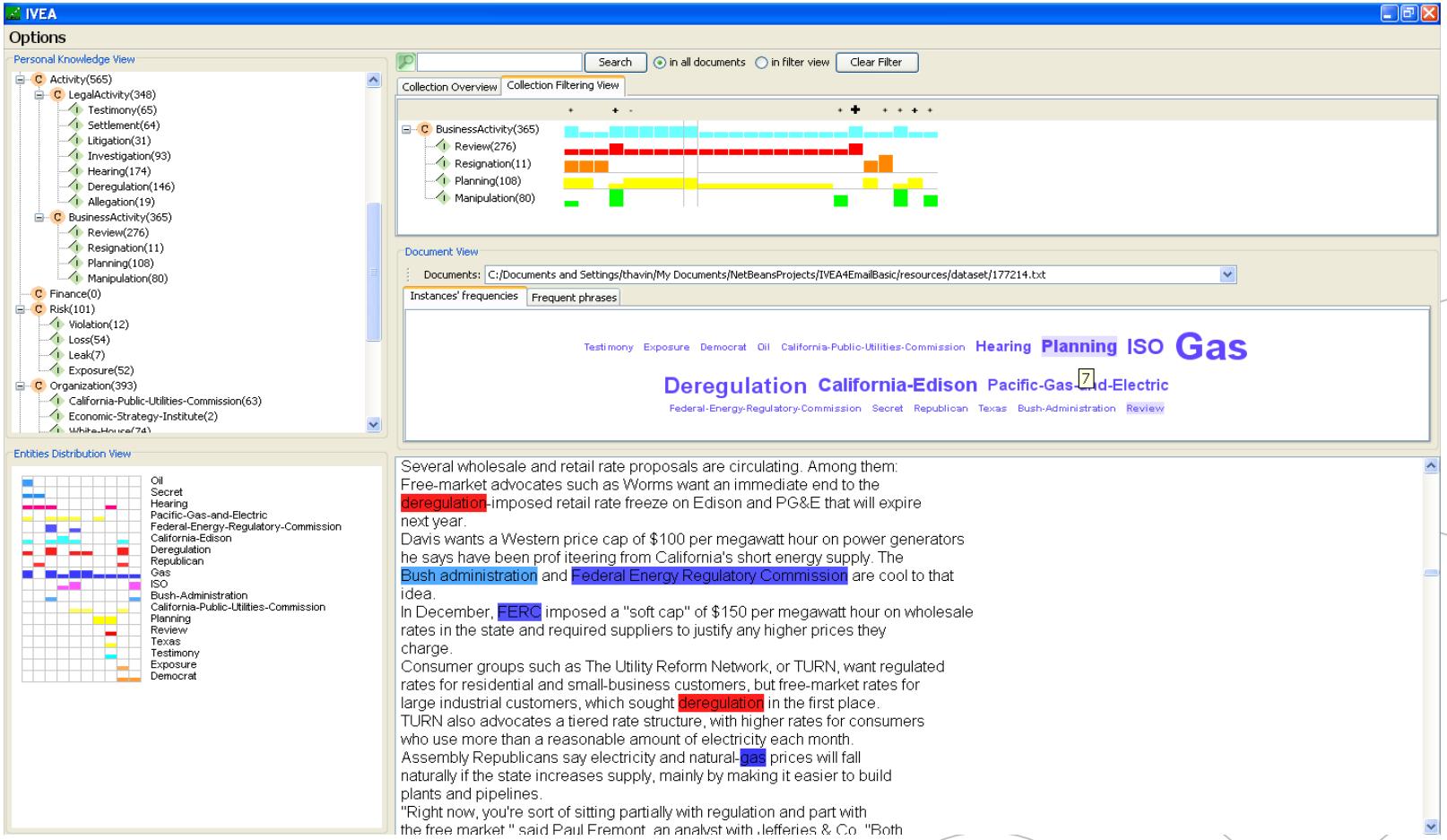
Proposed approach

- Advantages:
  - aligned with the users' interests
  - user-controlled
  - flexibility to explore at different levels of detail
  - dynamic ontology enrichment

→ more personalized to the users


Ontology Settings
Ontology
Ontology: ojects/IVEA4Email/resources/ontology/biz.owl
Load
LegalActivity
Testimony
Settlement
Litigation
Investigation
Hearing
Deregulation
Allegation
BusinessActivity
Review
Resignation
Planning
Manipulation
Finance
Risk
Violation
Loss
Leak
Exposure
Organization
California-Public-Utilities-Commission
Economic-Strategy-Institute
White-House
Bush-Administration
Federal-Energy-Regulatory-Commission
ISO
Los-Angeles-Times
Securities-and-Exchange-Commission
Chamber-of-Commerce
Clinton-Administration
Industry
Energy
Oil
Gas
Term variations
Save
federal energy regulatory commission
ferc
Done

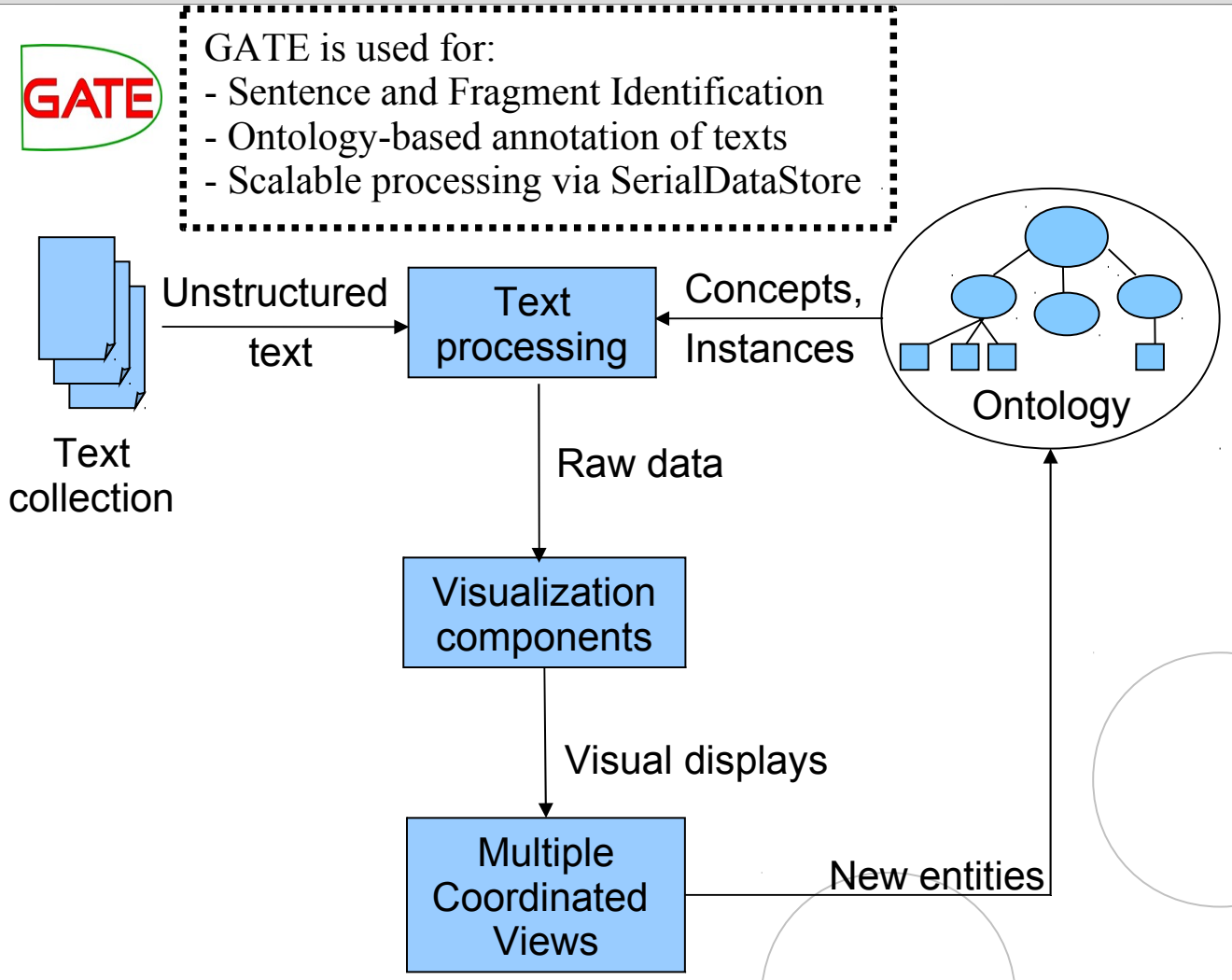
IVEA



http://smile.deri.ie/projects/ivea

sfi science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling networked knowledge.

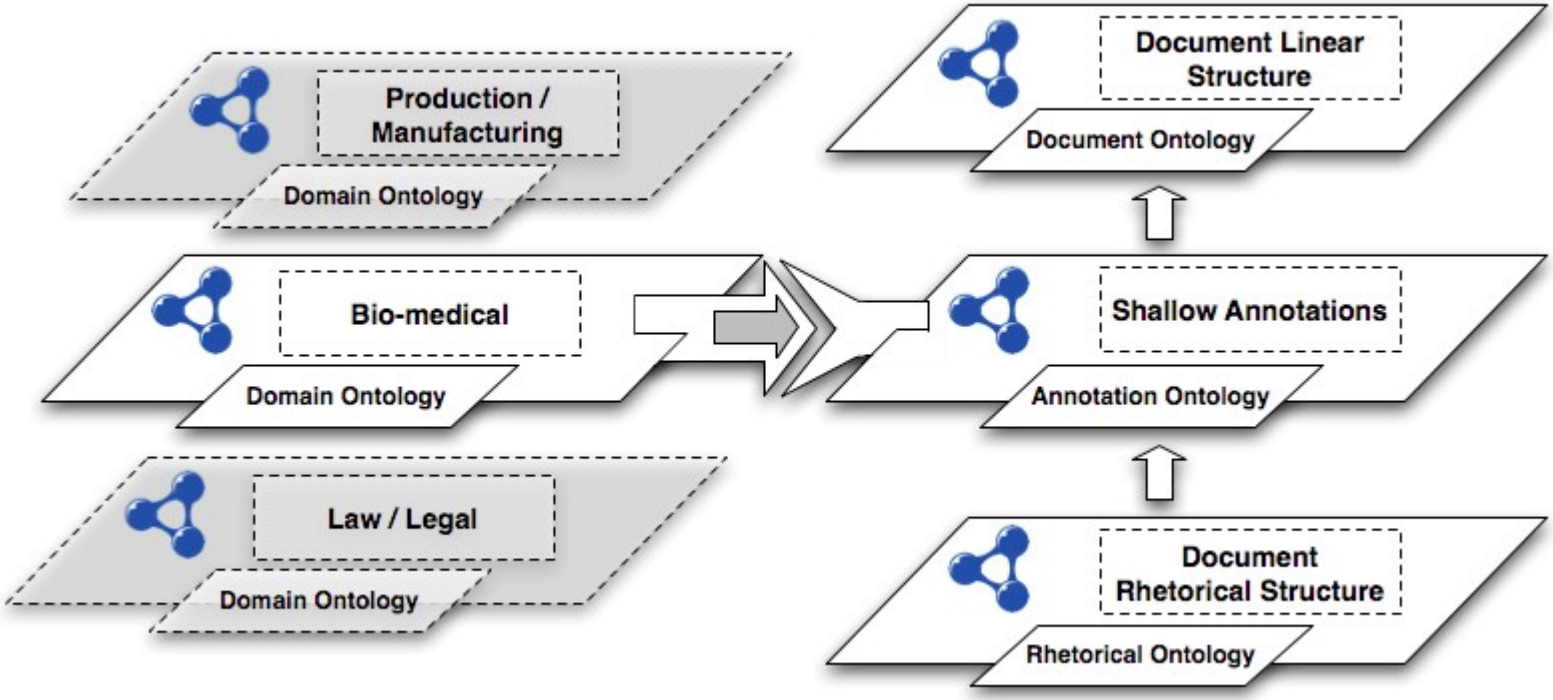
Role of GATE in IVEA



Embedding GATE in Semantic Web Applications II: SALT

science foundation ireland
fondúireacht eolaíochta éireann
NUI Galway
OÉ Gaillimh
Enabling networked knowledge.

SALT Ontologies



SALT - Goal

- **High-level**
  - **Automatic extraction of discourse knowledge items (i.e., claims, positions, arguments) from scientific publications**
- **Low-level**
  - **Rule-based extraction of rhetorical relations from the discourse, according to RST (Rhetorical Structure Theory)**

Approach

- **Empirical analysis**
- **Signalling discourse markers (cue-phrases)**
  - **Rhetorical relations: *however*, *although*, *but***
- **Two types of information**
  - **Discourse related information: type of rhetorical relations, roles of text spans**
  - **Algorithmic information: position, surrounding punctuation**
- **Result: Cue-phrases – Rhet. relations mapping**

Approach (cont.)

- **Extraction of rhetorical relations**
  - **GATE plugin**
  - **Empirical information encoded into fields for JAPE grammars**
    - **relation**
    - **whereToLink: A, B**
    - **statuses: SN, NS, NN**
    - **breakAction: NORMAL,**
    - **place: B, M, A**

Relation: Concession				
cue phrase	place	whereToLink	breakAction	statuses
. Although	B	A	NORMAL_THEN_COMMA	SN
, although	B	B	NORMAL_THEN_COMMA	NS
. However,	B	B	NORMAL	NS
, however	B	B	NORMAL	NS
, however,	M	B	NONE	NS
; however,	B	B	NORMAL_THEN_COMMA	NS
regardless	B	B	NORMAL_THEN_COMMA	NS
, though	B	B	NORMAL_THEN_COMMA	NS

Approach (cont.)

Relation: Concession				
cue phrase	**place**	**whereToLink**	**breakAction**	**statuses**
. Although	B	A	NORMAL_THEN_COMMA	SN
, although	B	B	NORMAL_THEN_COMMA	NS
. However,	B	B	NORMAL	NS
, however	B	B	NORMAL	NS
, however,	M	B	NONE	NS
; however,	B	B	NORMAL_THEN_COMMA	NS
regardless	B	B	NORMAL_THEN_COMMA	NS
, though	B	B	NORMAL_THEN_COMMA	NS

```
Rule:CuePhRule1 ({Token.string == ',' })
 ({Token.string == "while"}):cuePhrase1
                    ⟶
:cuePhrase1.CUE_PHRASE = {kind = 'cuePh',
   relation = "antithesis", place = B,
    breakAction = NORMAL_THEN_COMMA,
 statuses = NN, whereToLink = B, rule =
                "CuePhRule1"}
```


sfi science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh

Embedding GATE in Semantic Web Applications III: Semanta

Outline…

- Motivation
- Semantic Email
  - Email Action Items
  - Email Ad-hoc Workflows
- Semanta
  - Architecture
  - Email workflow support
  - Email workflow visualisation
  - Email workflow representation
  - Desktop Data Integration
- Evaluation & Discussion
- Conclusion

Motivation

Email remains the most popular means of *Electronic Communication*

- Asynchronous Communication
- Flexible, dynamic nature

Email is also a *Virtual Workplace*

- *Collaborative Environment*
- Knowledge creation, management and sharing

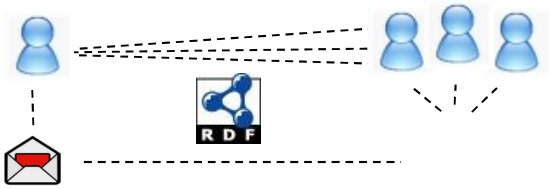
Lacks clear structure & real semantics → Email Problems

- Email Tracking
- Email Classification
- Email Retrieval
- *Email Overload*

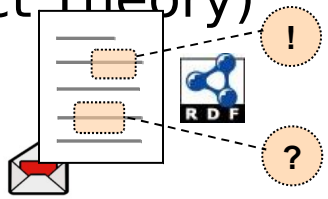
Semantic Email

Annotated Email

i. *Thread metadata -* Email Sequence, Social, Temporal Metadata

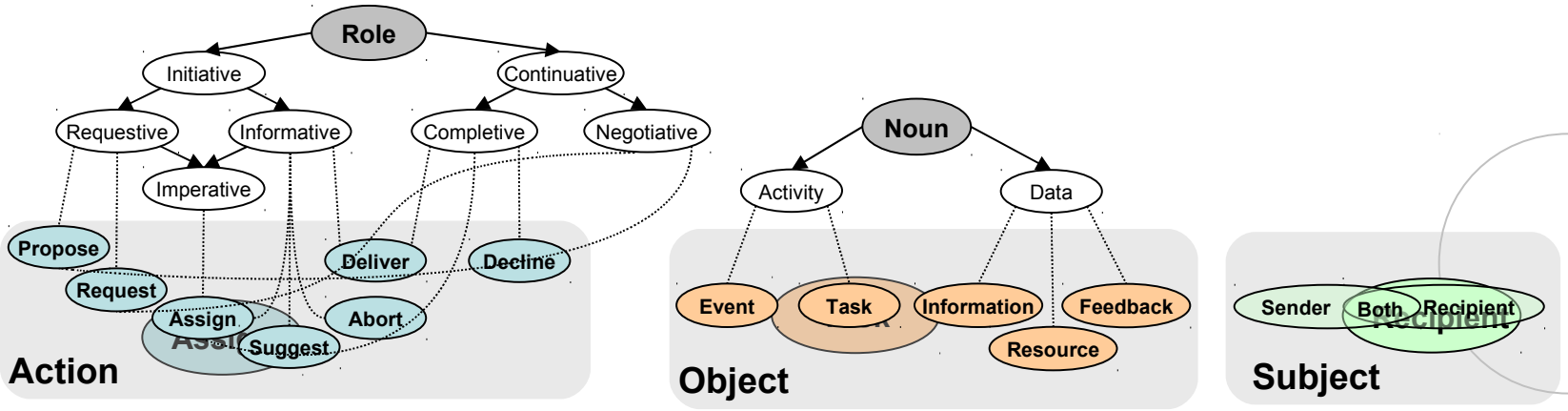


ii. *Content metadata –* Action Items in written dialogue (based on Speech Act Theory)



Email Action Items

Speech Act Model: [***Action***, ***Object***, ***Subject***]



Example *"...Please make sure you have the document ready!.."*

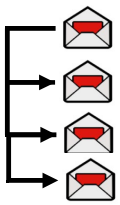
sfi science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling **networked** knowledge.

Email Workflows

Email Conversations carry out concurrent, implicit, well-formed *Workflows*

Email Action Item = Start/Continuation of a Workflow

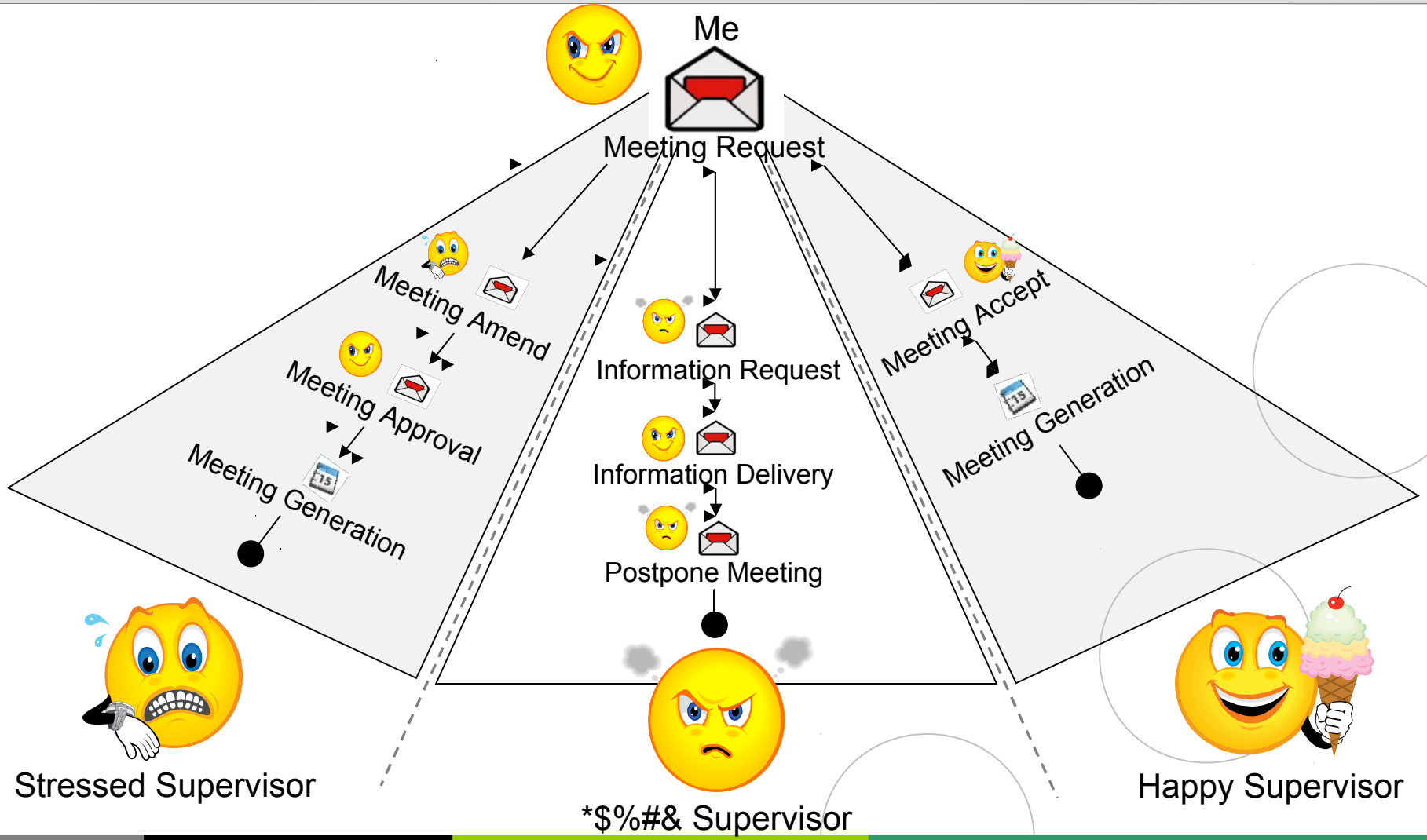
Example:



- ***Request*** Meeting
- ***Amend*** Meeting properties
- ***Approve*** Meeting
- ***Announce*** Meeting

•*Workflow Artefacts* –Events, Tasks, People, Projects…

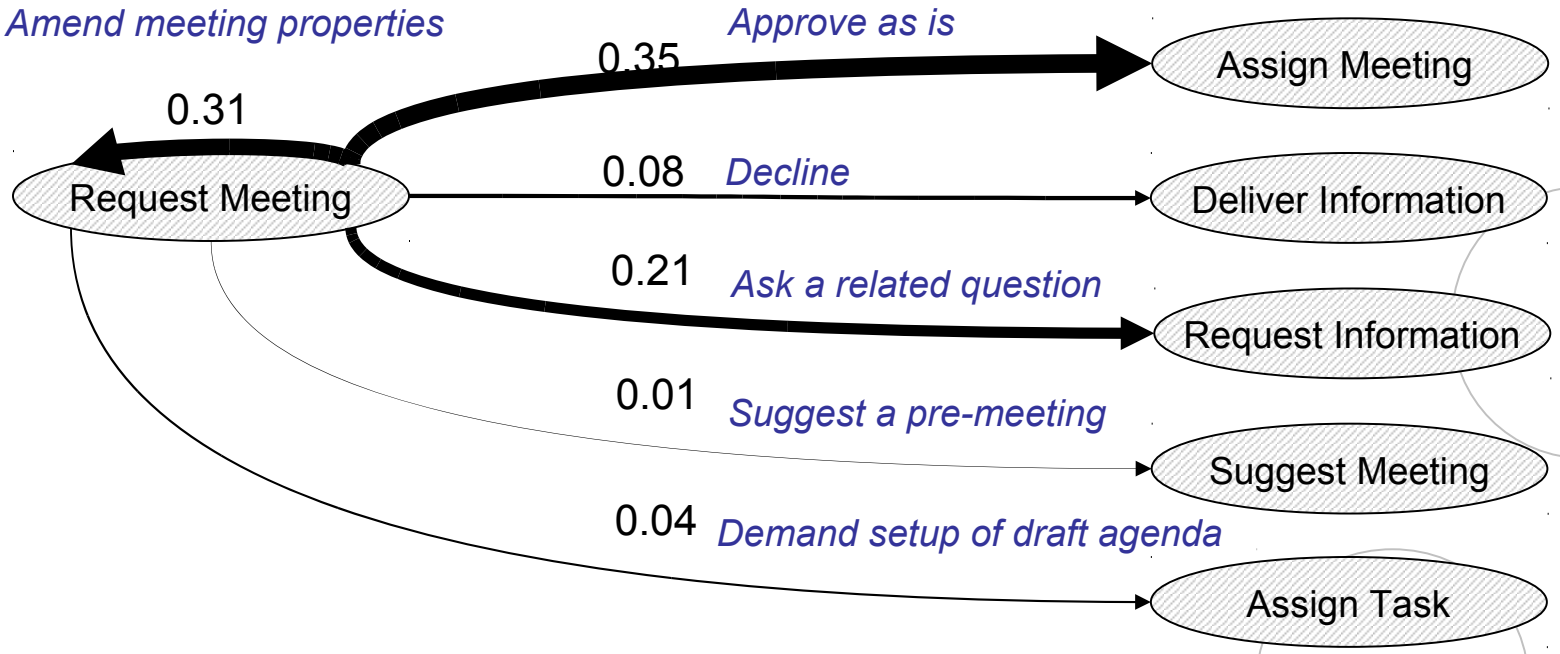
Email Ad-Hoc Workflows



science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling networked knowledge.

Modelling Email Ad-Hoc Workflows

- There exist ***trends***

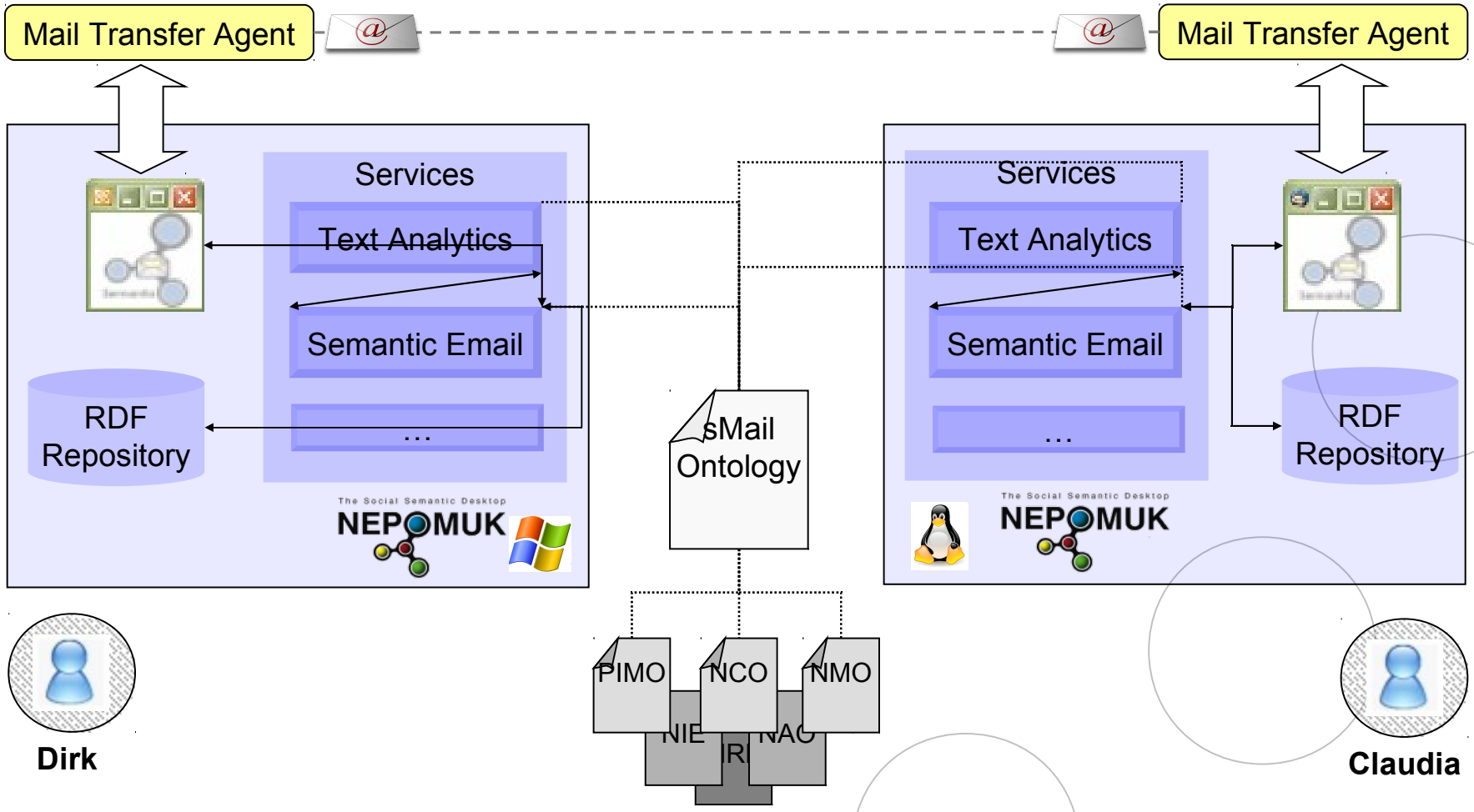


- Provide support for the *most-likely* Action Item reactions
- Leave open the option for *any other* reaction

Semanta - Features

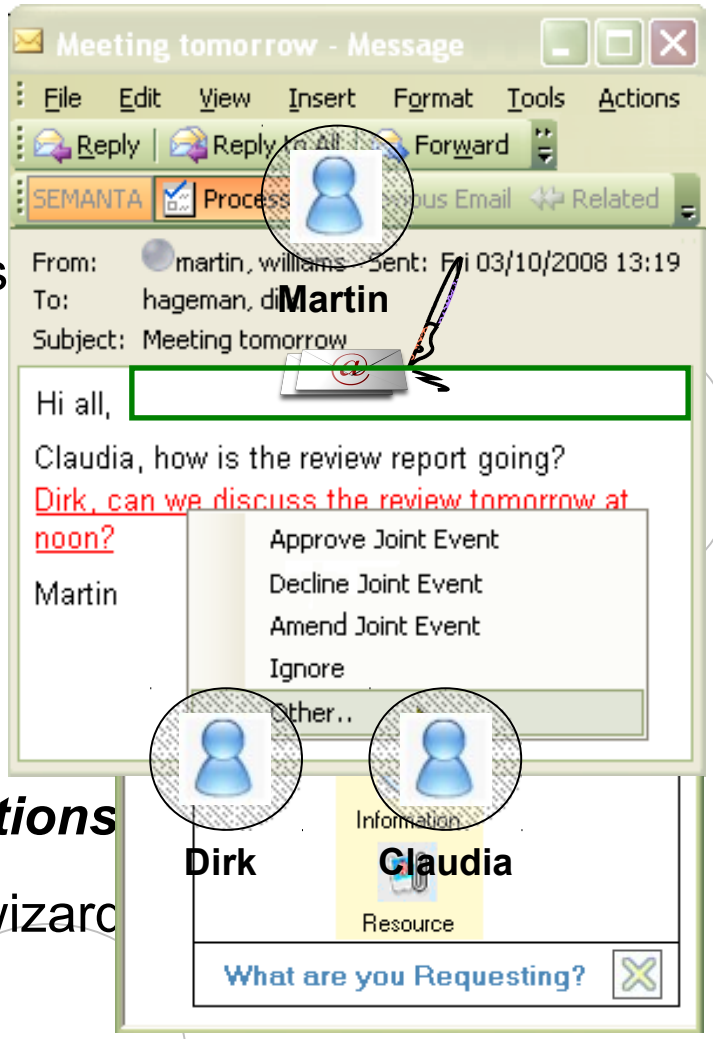
- Semi-automatic Action Items *Detection*
- Supporting Email Action Item *Handling*
- Email Action Item (Workflow) *Tracking*
- Email Workflow *Visualisation*
- Email-generated Event/Task *Recognition*
- Email attachment *Reminders*
- Email←→ Desktop *Knowledge Integration*

Semanta - Architecture



Email Workflow Support - Example

- Martin writes an email to Dirk and Claudia
- Action Items ***extracted*** from content
- ***Annotation Wizard*** - add/modify annotations
- Semantic Email is sent…
- …Dirk receives the email
- Dirk's Action Items are ***highlighted***
- Semanta supports Dirk with ***most likely reactions***
- …and any ***other*** reaction via the annotation wizard

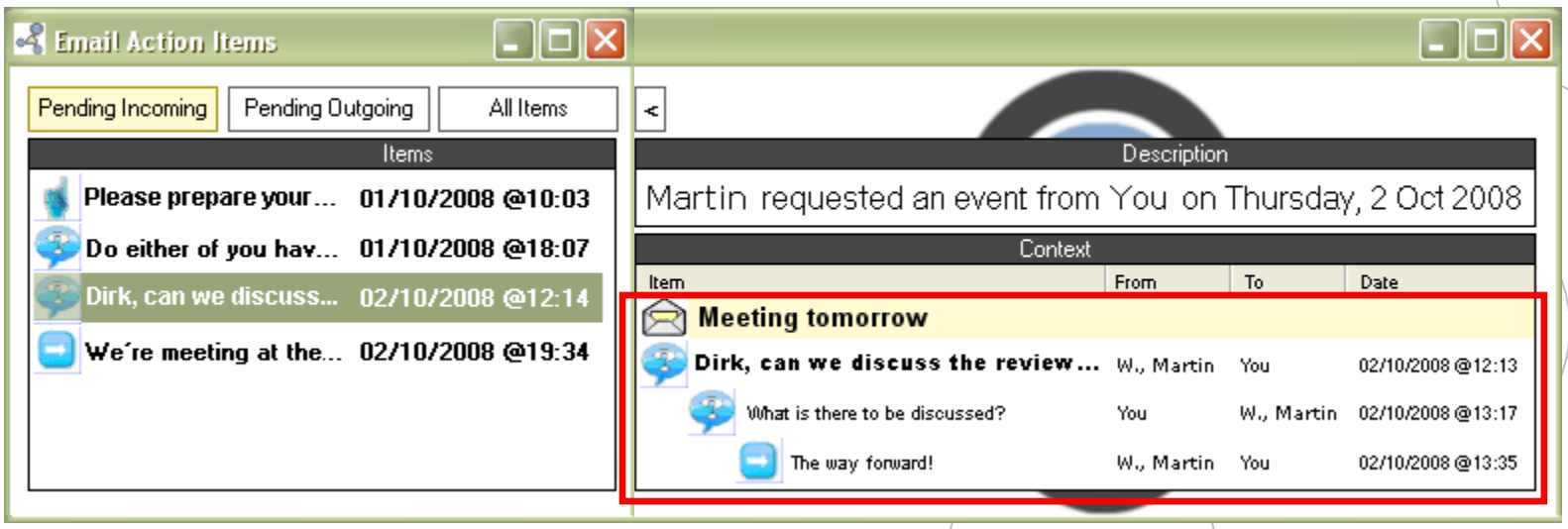


Email Workflow Visualisation

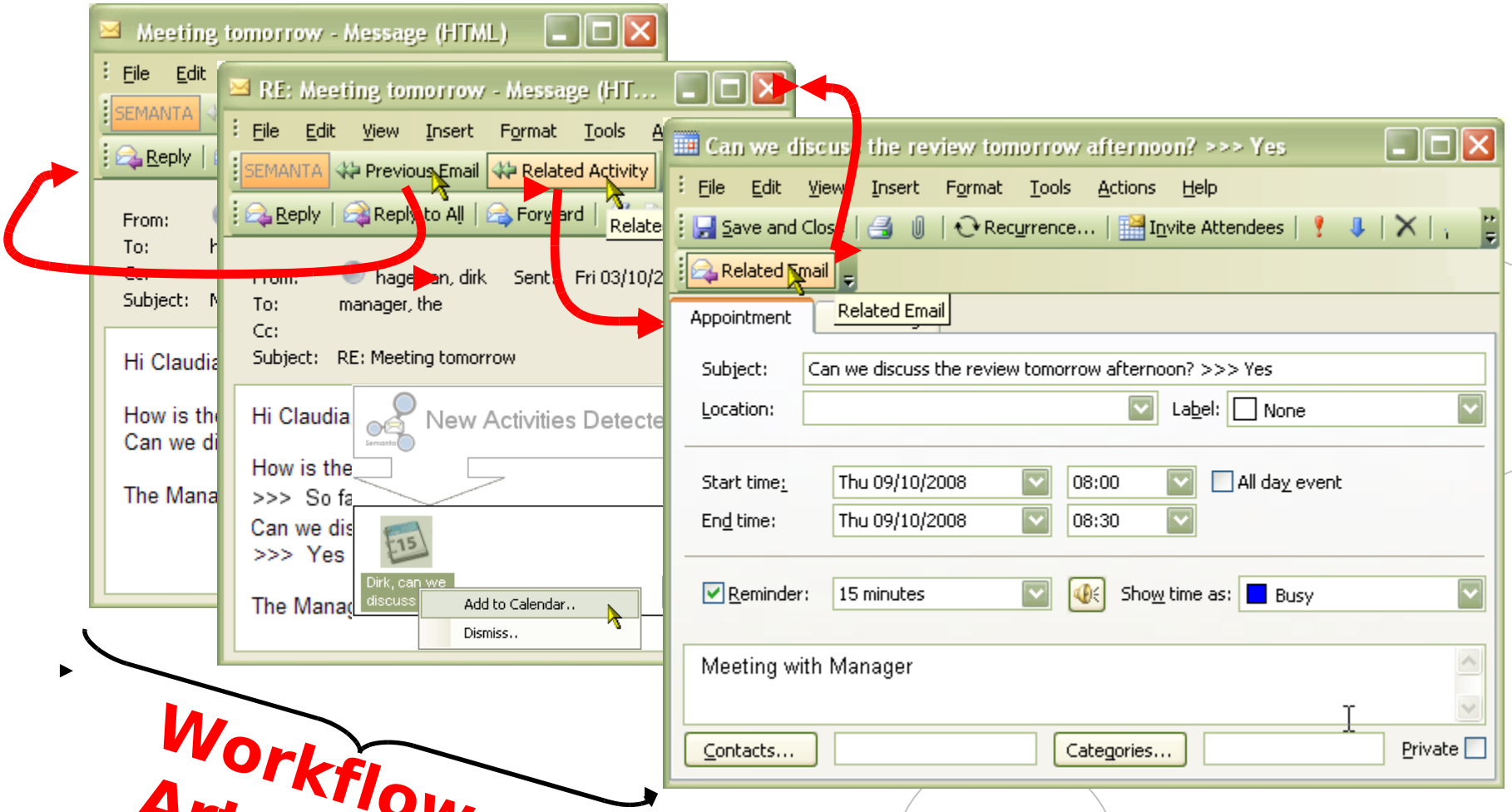
- Pending Incoming Action Items :- Personal Email Todo's
- Pending Outgoing Action Items :- Sent requests which remain unanswered
- All Items :- All sent & received items

WORKFLOW

Individual Items in all views can be shown in their CONTEXT

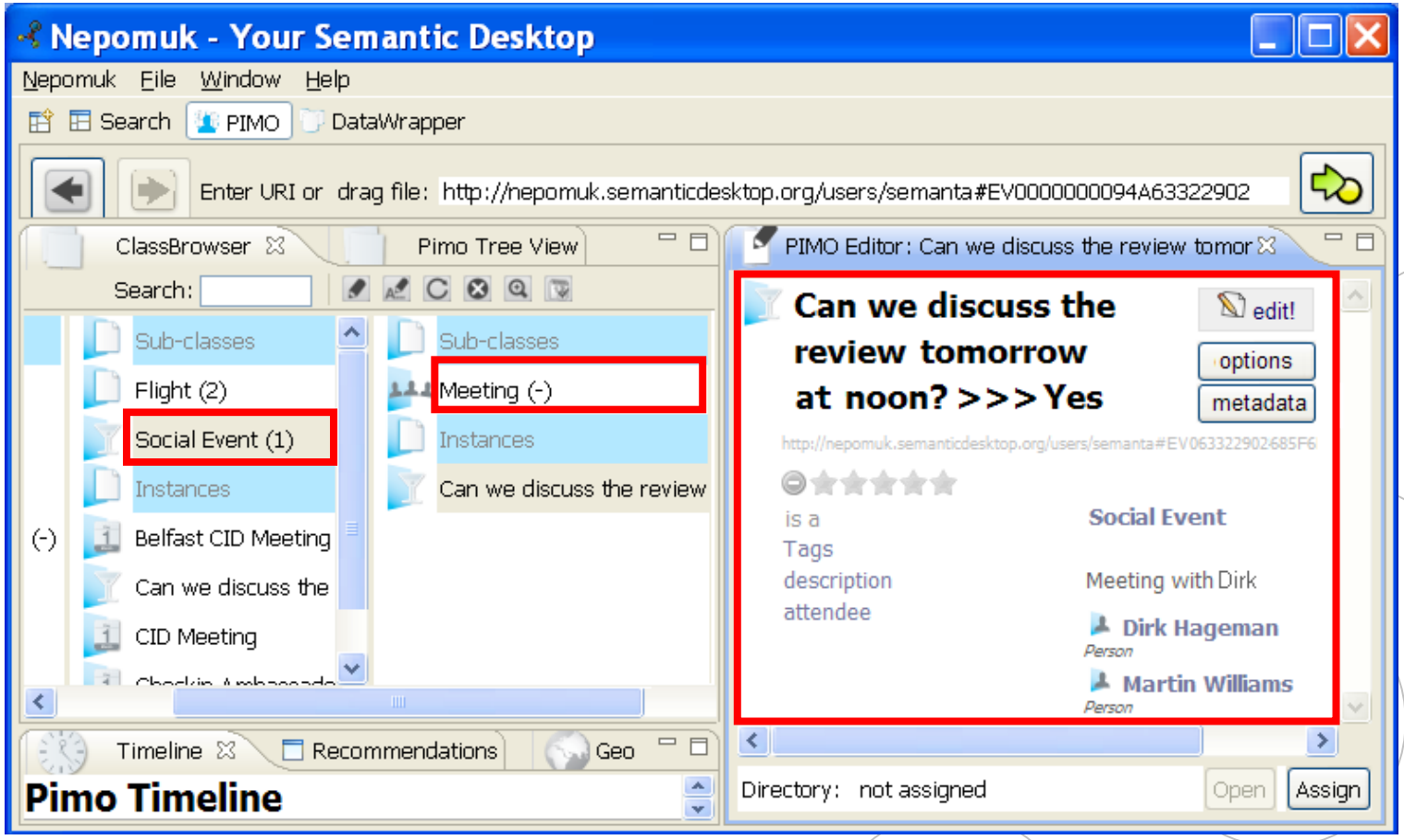


Email Workflow Representation



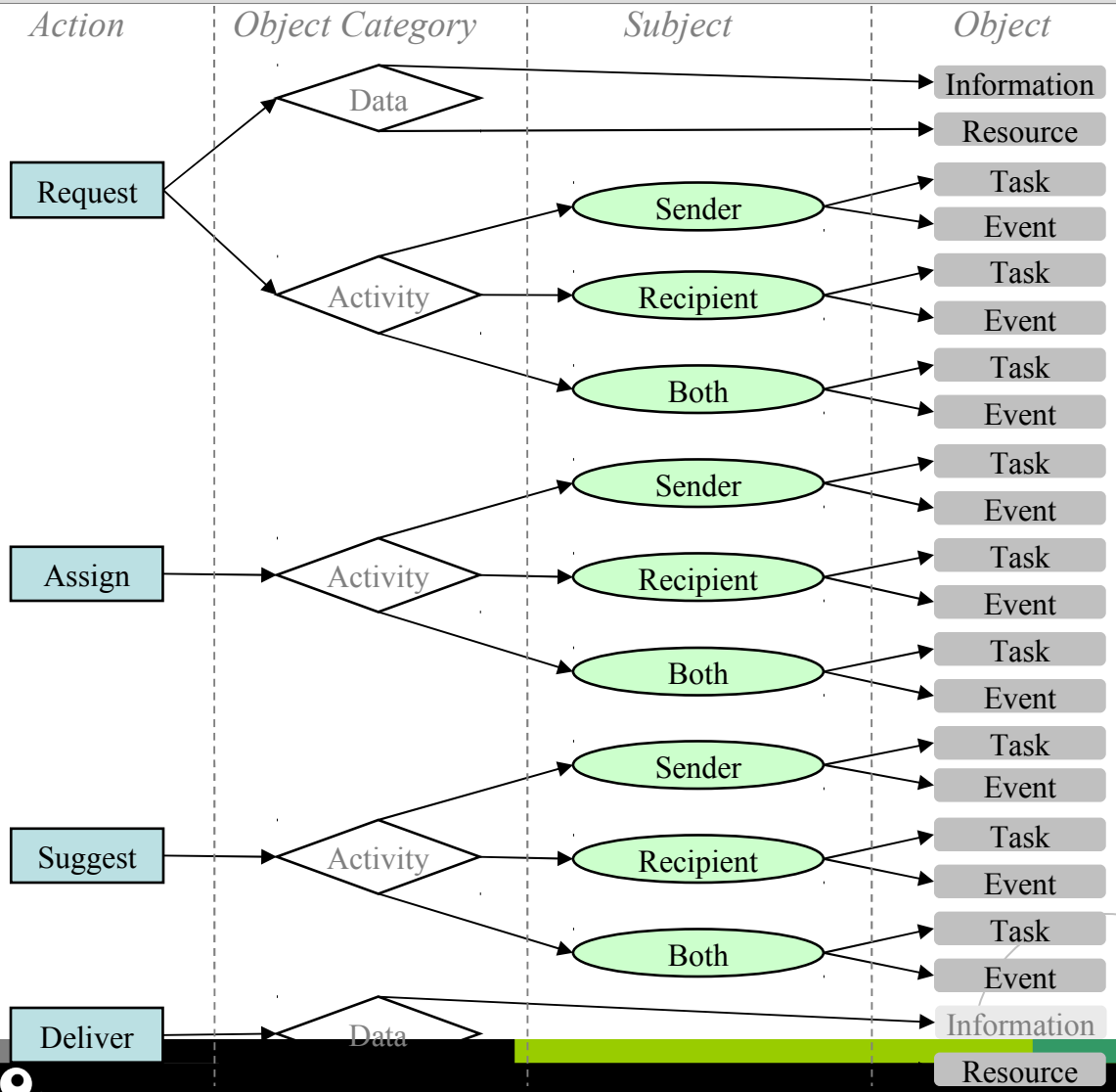
sfi science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling **networked** knowledge.

Desktop Data Integration



sfi science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling networked knowledge.

21 Classes for Email Speech Act Classification

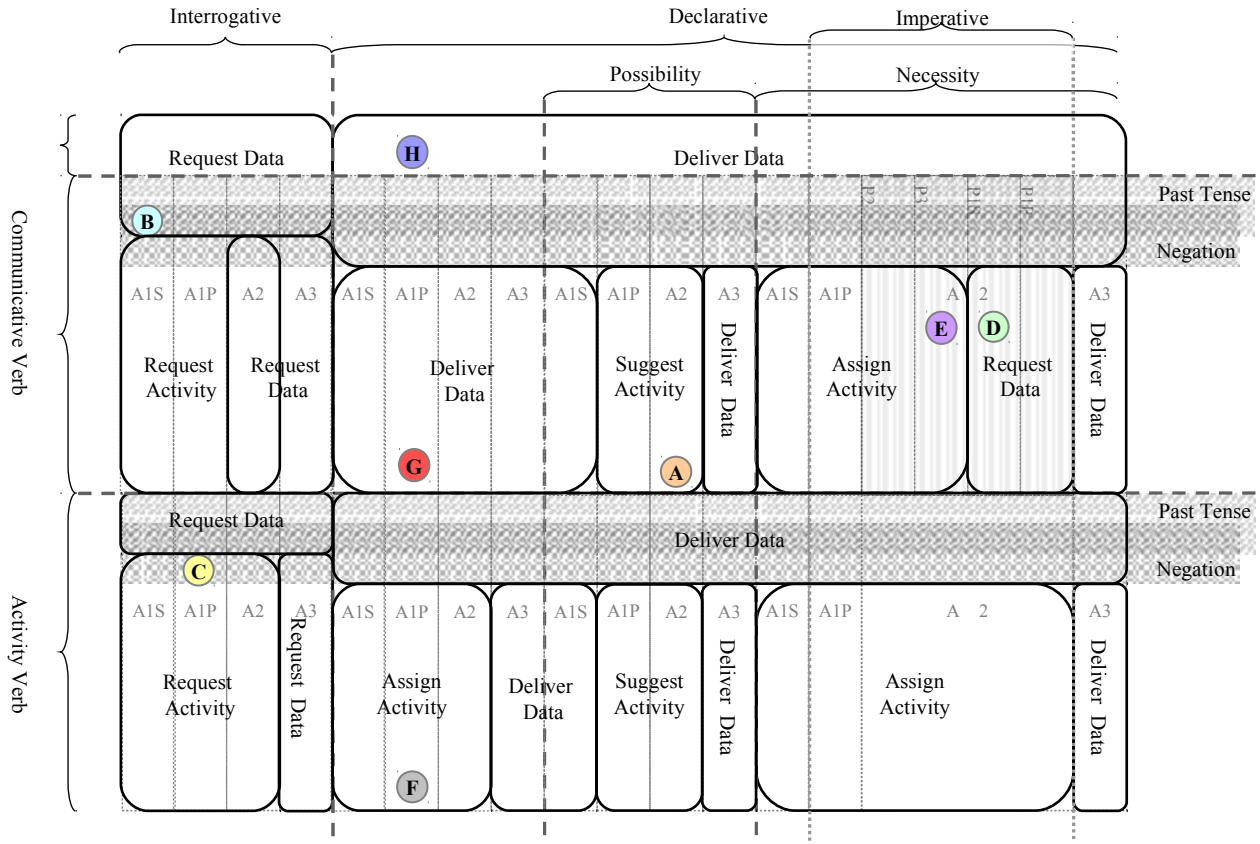


Request information from recipient(s)
Request a file from recipient(s)
Request permission for a personal task
Request permission for a personal event
Request a task from recipient(s)
Request the recipient(s)'s attendance to an event
Request a joint task from recipient(s)
Request joint event participation from recipient(s)
Commit yourself to a task and inform recipient
Commit yourself to an event
Assign a task to recipient(s)
Assign an event to recipient(s)
Announce a joint task to recipient(s)
Announce a joint event to recipient(s)
Volunteer for a task performance
Volunteer for an event attendance
Suggest a task to recipient(s)
Suggest the recipient(s)'s attendance to an event
Suggest a joint task to recipient(s)
Sugget joint event participation to recipient(s)
Deliver information to recipient(s)

Classification Model

- **Classification Task – elicit e-mail speech acts from e-mail content**
- **Classification Model maps a text clause into exactly one of the 21 classes**
- **Text classification is based on:**
  - **Sentence Form (Interrogative, Declaritve, Imperative)**
  - **Verbal Modality (Possibility, Necessity)**
  - **Verb Type (Activity, Communicative, Other)**
  - **Semantic Role (Agent, Patient, 1st/2nd/3rd Person Singular/Plural)**
  - **Negation (Presence of)**
  - **Grammatical Tense (Past and Non-Past)**

Graphical Representation of Classification Model

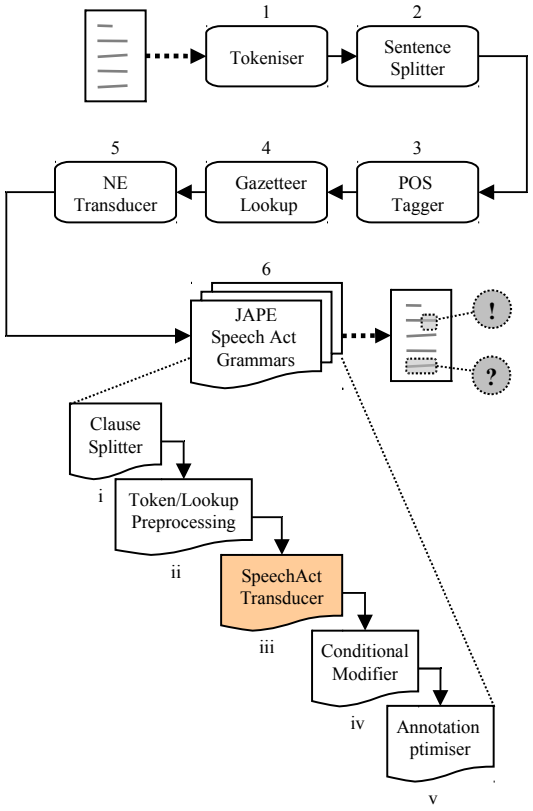


Examples

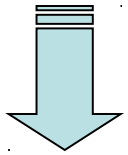
A	You should forward it to me.
B	Haven't I sent you the file?
C	Didn't we need to discuss today?
D	You still have to send me the info!
E	You must email them the data.
F	We are attending the meeting.
G	We are sending you the files.
H	We are happy.

Gate Implementation

The Pipeline

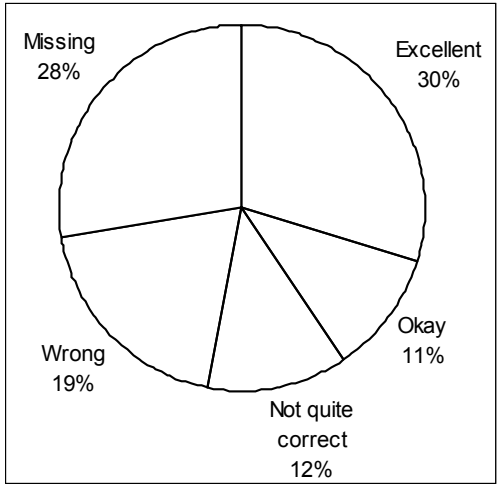


Speech Act Transducer



- **Performs bulk of pattern matching**
- **Intermediate annotations matched to 1 class**
- **Consists of 58 rules**
- **Rules fire within 14 different phases**
- **Text matched in the initial phases not considered later**

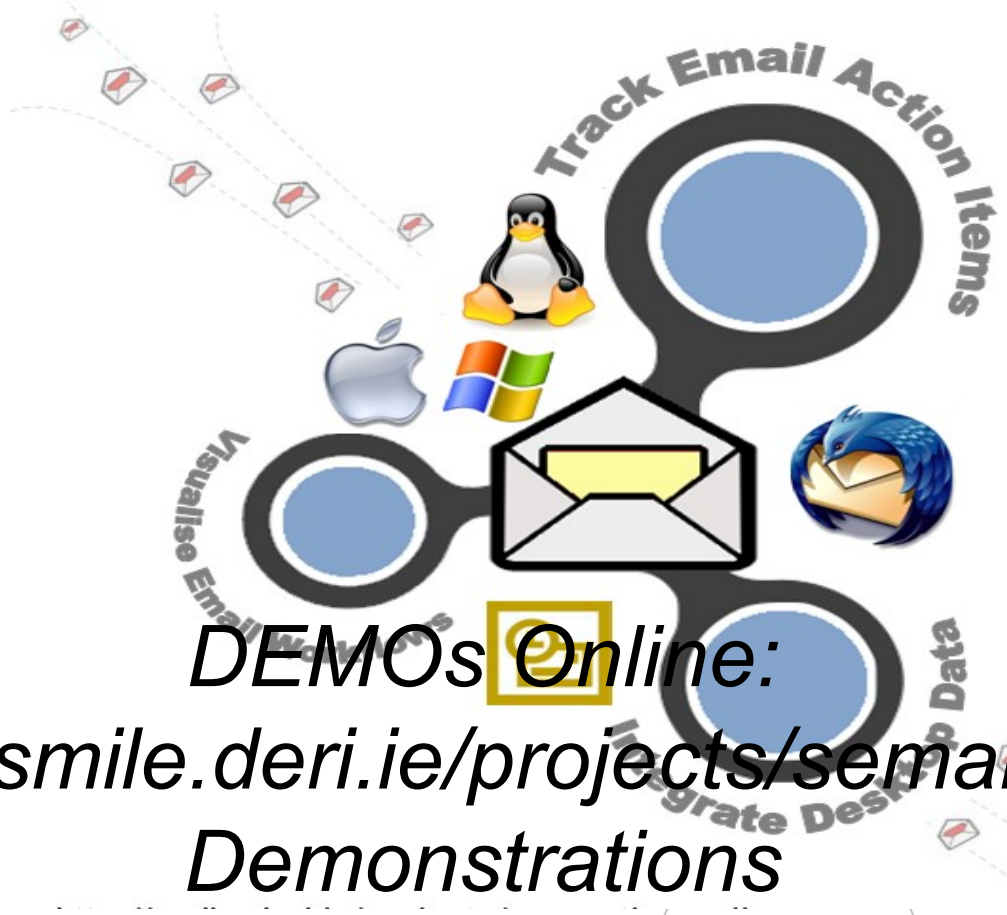
Evaluation – Some Results



- **12 E-mail users rated results of automatic classification**
- **Ratings: Excellent, Okay, Not quite correct, Wrong**
- **Evaluators asked to highlight missing e-mail Action Items**
- ***F*-measure of 0.58 (Precision 0.56, Recall 0.60)**
- **Earlier human inter-annotator agreement: 0.811**
- **Conclusion: Not reliable for automatic classification**
- **Result: Employed semi-automatic to provide suggestions**

Semanta – Semantic Email in Action

titute



DEMOs Online:
http://smile.deri.ie/projects/semanta/#Demonstrations

sfi science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling **networked** knowledge.

Interfacing to Semantic Web Technologies with GATE : CLOnE and ROA

science foundation ireland
fondúireacht eolaíochta éireann
NUI Galway
OÉ Gaillimh
Enabling networked knowledge.

Motivation

Problem

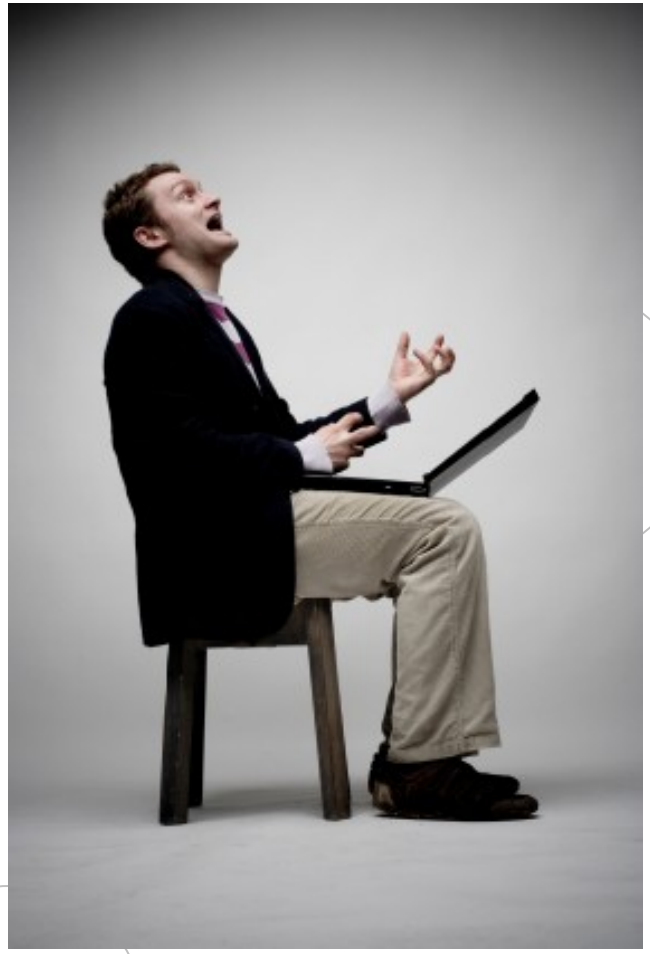
- Domain experts are not Ontology Engineers !!

Goal

- allows a quick easy first draft of a complex Ontology
- creation of small to medium sized Ontologies by novice users

Support

- large percentage of an initial Ontology would naturally consist of taxonomic relations and simple properties/relations



Introduction

Controlled Language for Ontology Engineering

What is Controlled Language?

What is the Problem with Controlled Language?

- Habitability Problem
- Learning Curve

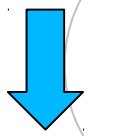
Our Contribution

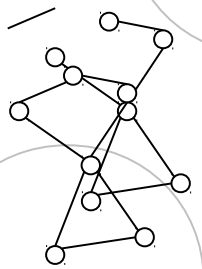
- Round Trip Ontology Authoring : Combining Controlled Language with Language Generation
- Empirical Evidence

Goals: Controlled Language

Provide controlled language for basic ontology-editing functions

- easy to learn from examples and simple rules
- relatively easy to deploy (Java, GATE)
- unambiguous
- compact (e.g., create many classes or instances with one sentence)
- natural but grammatically lax



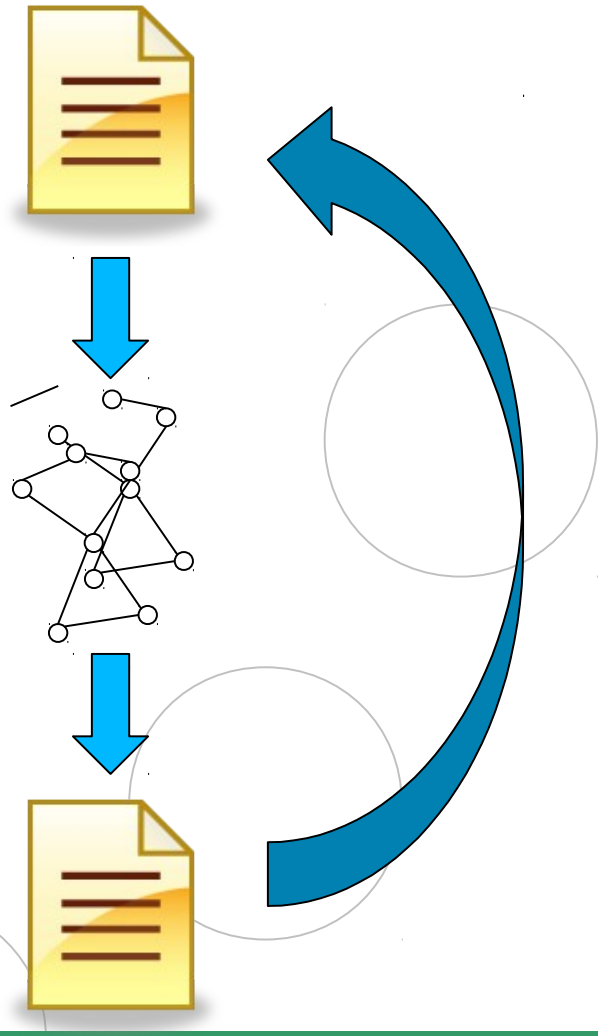


Outcome: →

CLOnE: Controlled Language for Ontology Editing.

Goals: RoundTrip Ontology Authoring (ROA)

- Improve on existing user-friendliness
- Ease CLOnE learning experience using Natural Language Generation (NLG)
- Simple Ontology summarisation via Natural Language Generation



Research Questions– RoundTrip Ontology Authoring (ROA)

Can NLG effectively substitute for CL style guides?

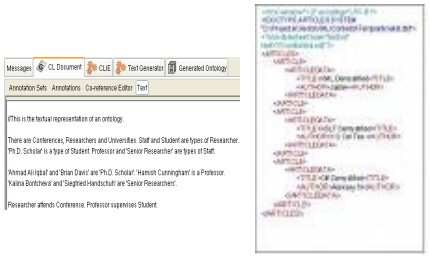
Can NLG help ease the habitability problem?

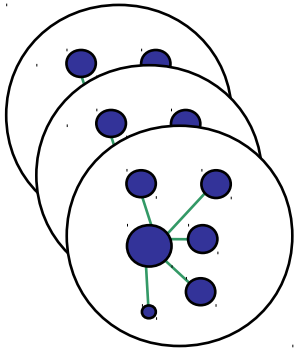
Can NLG improve on previous evaluation results?

What is NLG?

Round Trip Ontology Authoring (ROA)

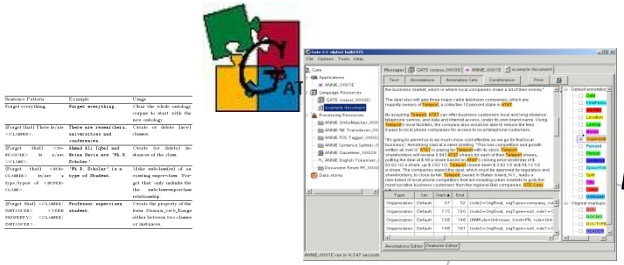
Natural Language Generation (NLG)





Ontology

Controlled Language for Ontology Engineering (CLOnE)

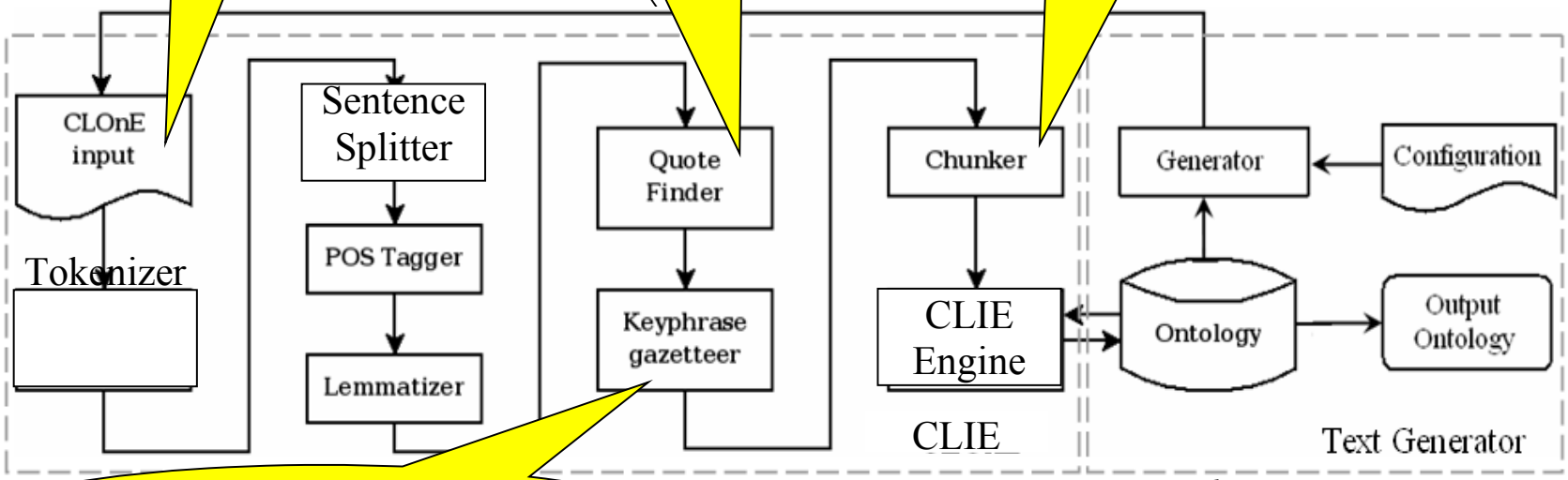


Controlled Language for IE (CLIE)

Examples of CLOnE in ROA

Example	Pattern	Usage
Forget everything	`Forget everything.`	Clean Ontology
There are researcher, universities and conferences.	`There is/are <CLASSES>.`	Create classes.
Brian Davis and Simon Scerri are 'Ph.D. Scholar'.	`<INSTANCES> is/are <CLASS>.`	Create instances of the class.
'Ph.D. Scholar' is a type of student.	`<SUB-CLASSES> is a type of <CLASS>.`	Make sub-classes.
Professor supervises student.	`<CLASSES/INSTANCES> <VERB PROPERTY> <CLASSES/INSTANCES>`	Create the property

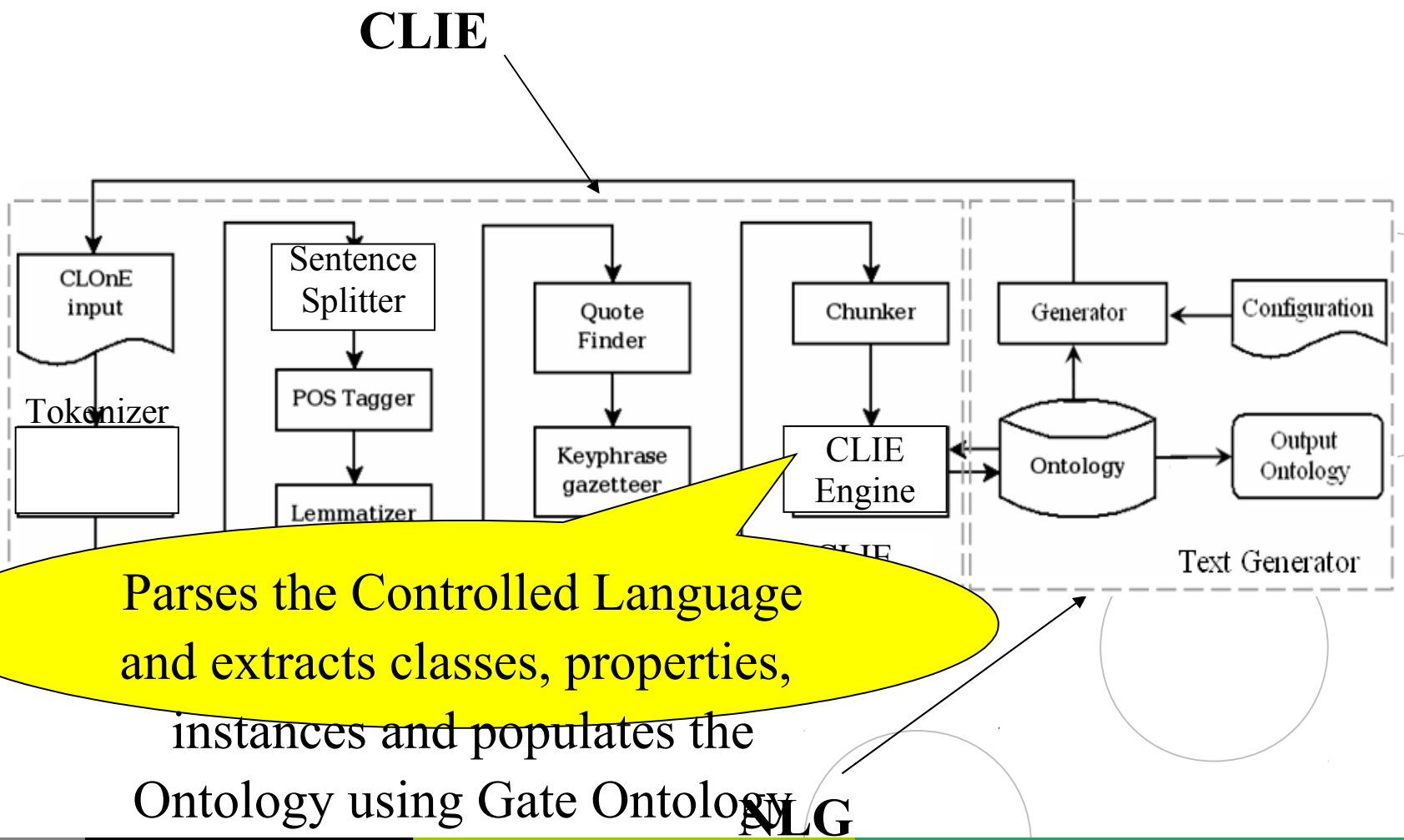
ROA – Processing Pipeline



Dictionary for reserved words, e.g. "there are", "is a"

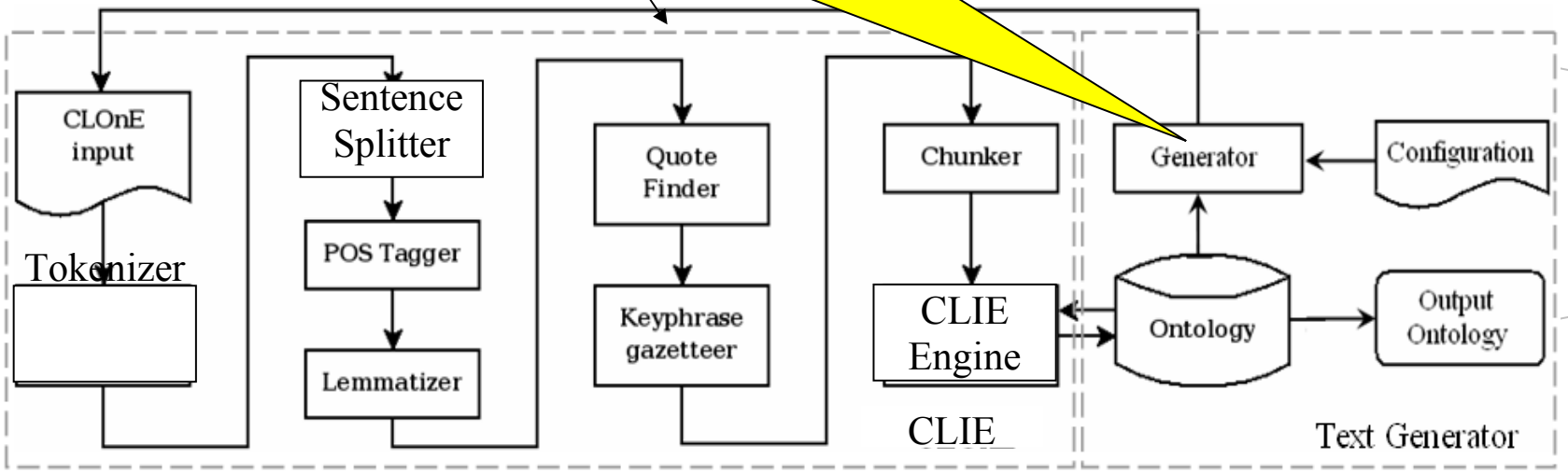
NLG

ROA – Processing Pipeline



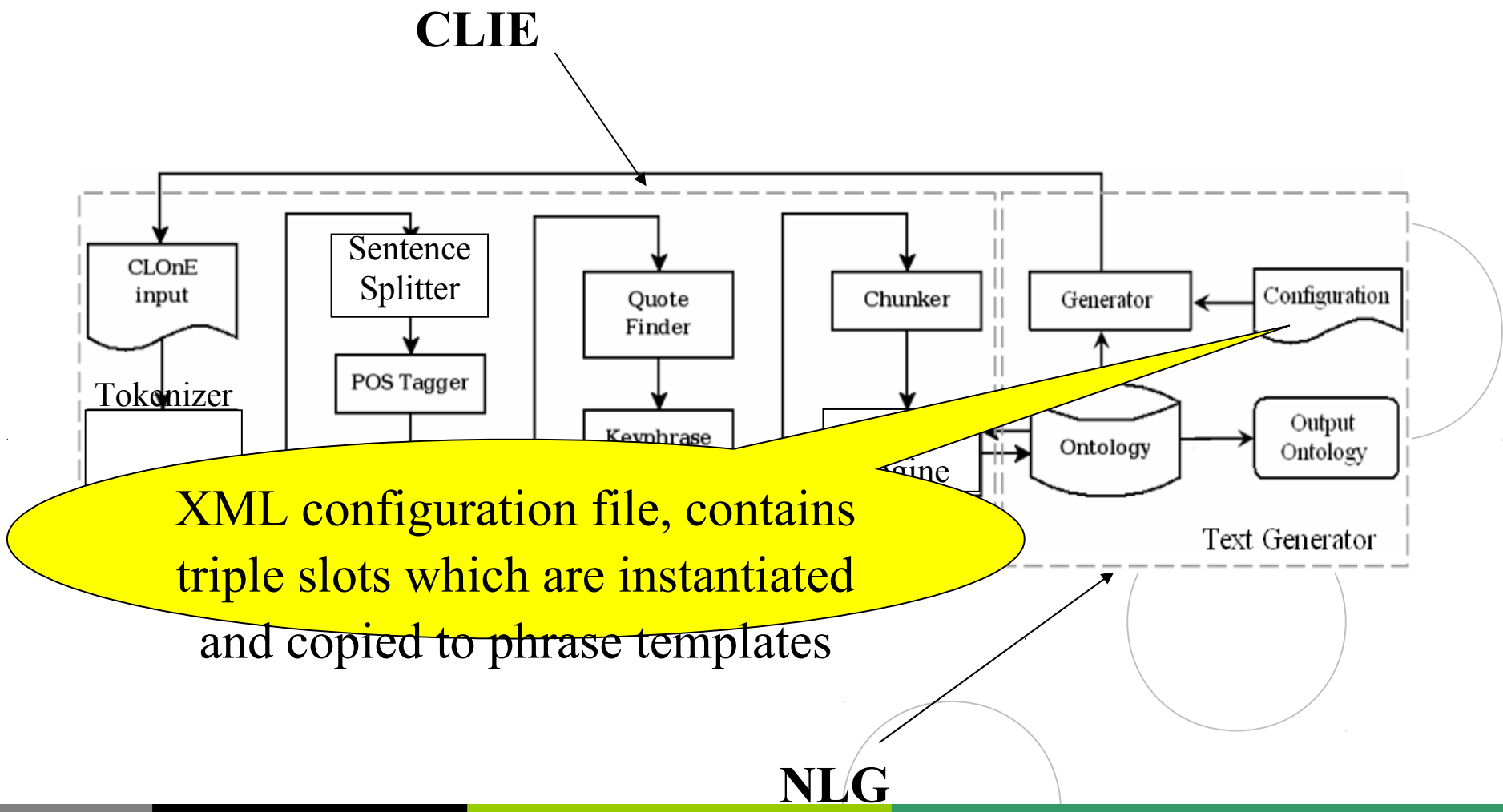
ROA – Processing Pipeline

Text generator flattens Ontology into Triples matches to configuration file template slots

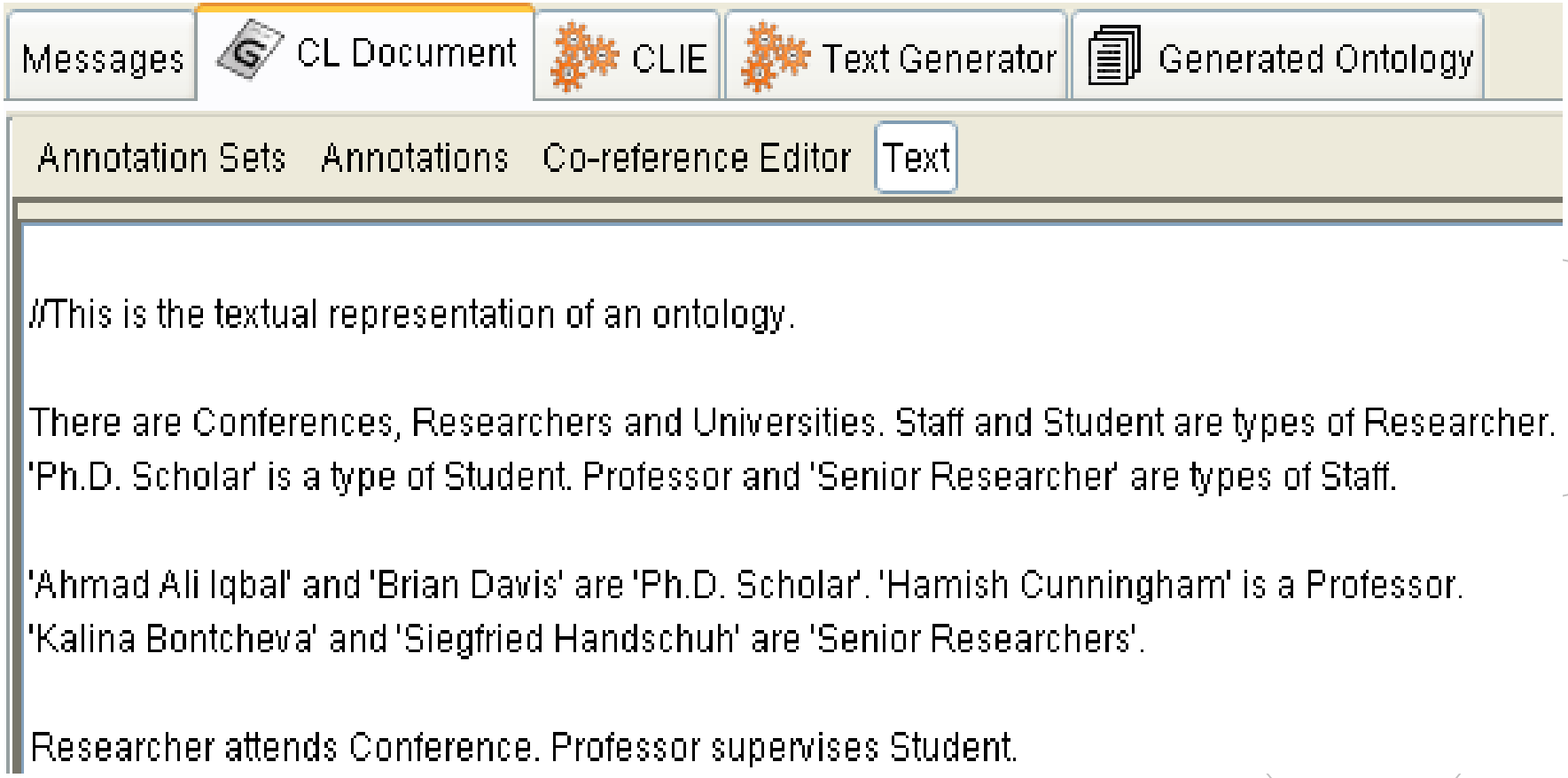


NLG

ROA – Processing Pipeline



ROA: Generator Output

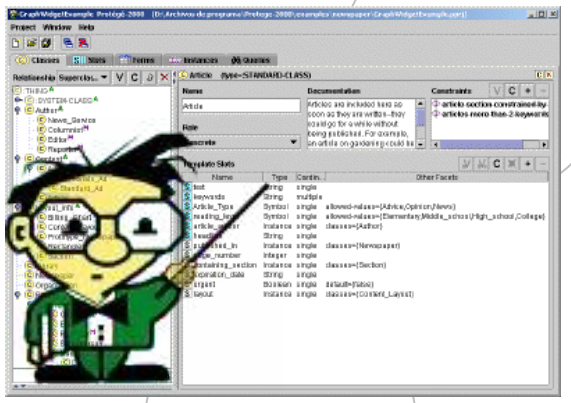


Evaluation: Compare with Protégé

Evaluation: Compare ROA with Protégé

- Protégé is the standard tool for ontology authoring
- Previous work compared CLOnE with Protégé.
  - compare ROA with CloNE – it was necessary to include Protégé in order to repeat the experiment
- Note: We make no claims that Protégé would be replaced by ROA !

VS



Evaluation Setup

Methodology based on previous Controlled Language (ClOnE) evaluation

- Ensured fair comparison between ROA and CLOnE
- Using the System Usability Scale (SUS)[1]

CLOnE reference manual and example withheld!

- Substituted with text generator

[1] http://en.wikipedia.org/wiki/System_Usability_Scale

Evaluation

Preparation

- Pre-test questionnaire to let users rate their own knowledge of ontologies and CLs
- Short manual on ontologies and both tools

Sample Quality

- Bigger sample size !
- Tighter control over bias!
- Consistent evaluation values between Researcher and Industry groups

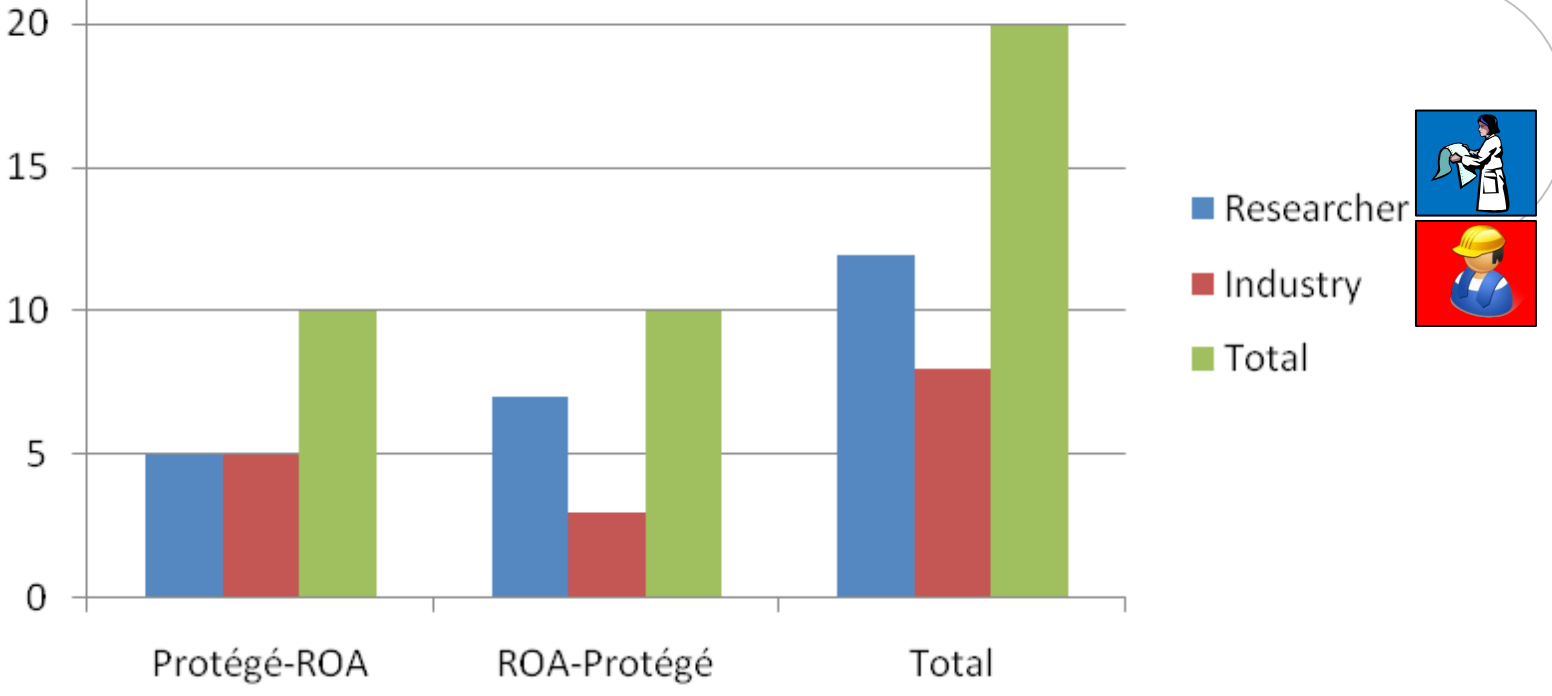
Evaluation Type:

- Repeated measures, tasked-based evaluation

Evaluation: Tasks

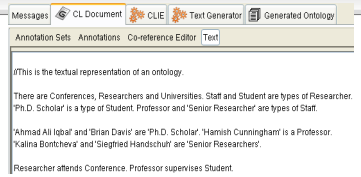
Two progressive lists of 6 simple tasks, task list A & B

- Group 1: **A**: Protégé → **B**: ROA
- Group 2: **A**: ROA → **B**: Protégé

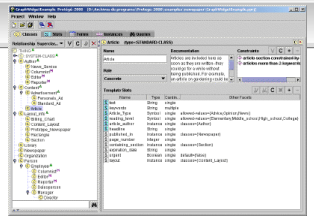


Group 1 Group 2

Evaluation results: ROA vs. Protégé



ROA

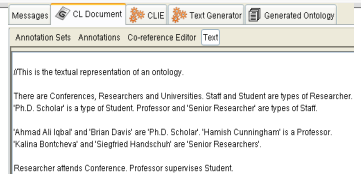


Protégé

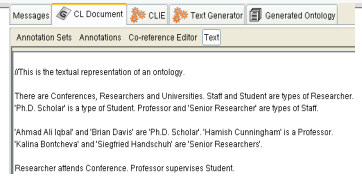
High SUS score	☑	■ Low SUS score
Suitable for both Tasks		■ Suitable for both Tasks
Industrial Users	☑	■ Industrial Users
Non-Expert Users	☑	■ Non-Expert Users
Less Time	☑	■ More Time

science foundation ireland
fondúireacht eolaíochta éireann
NUI Galway
OÉ Gaillimh
Enabling **networked** knowledge.

Evaluation results: ROA vs. ClOnE?



ROA



CLOnE

Maintained Interest	☑	■ Waning Interest
No Manual needed	☑	■ Manual needed
Less Time	☑	■ More Time
Higher Satisfaction	☑	■ Lower Satisfaction

Overall new evaluation allows to make claims over the entire population.

science foundation ireland
fondúireacht eolaíochta éireann
NUI Galway
OÉ Gaillimh
Enabling **networked** knowledge.

Conclusion and Future Work

Summary

- IVEA
- SALT
- Semanta

Ongoing and Future work

- Scalable processing in IVEA with GATE
- ML for SALT
- ML for Semanta

References

IVEA

http://smile.deri.ie/projects/ivea

Contact Point: VinhTuan Thai

SALT

http://salt.semanticauthoring.org/

Contact Point: Tudor Groza

Semanta

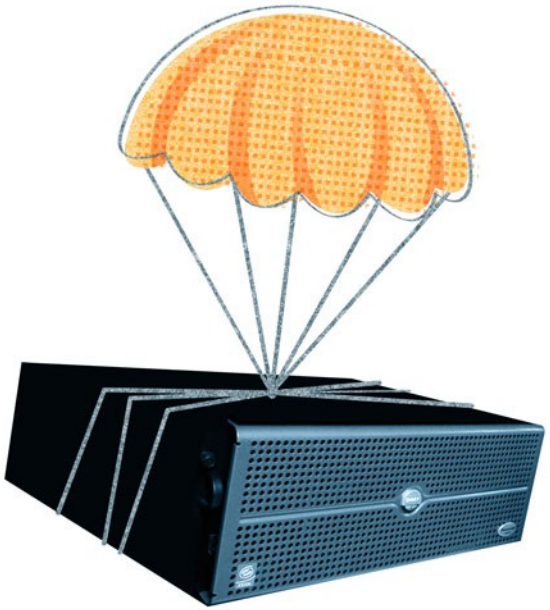
http://smile.deri.ie/projects/semanta/

Contact Point: Simon Scerri

CLOnE and ROA

http://smile.deri.ie/evaluation/2008/ROA

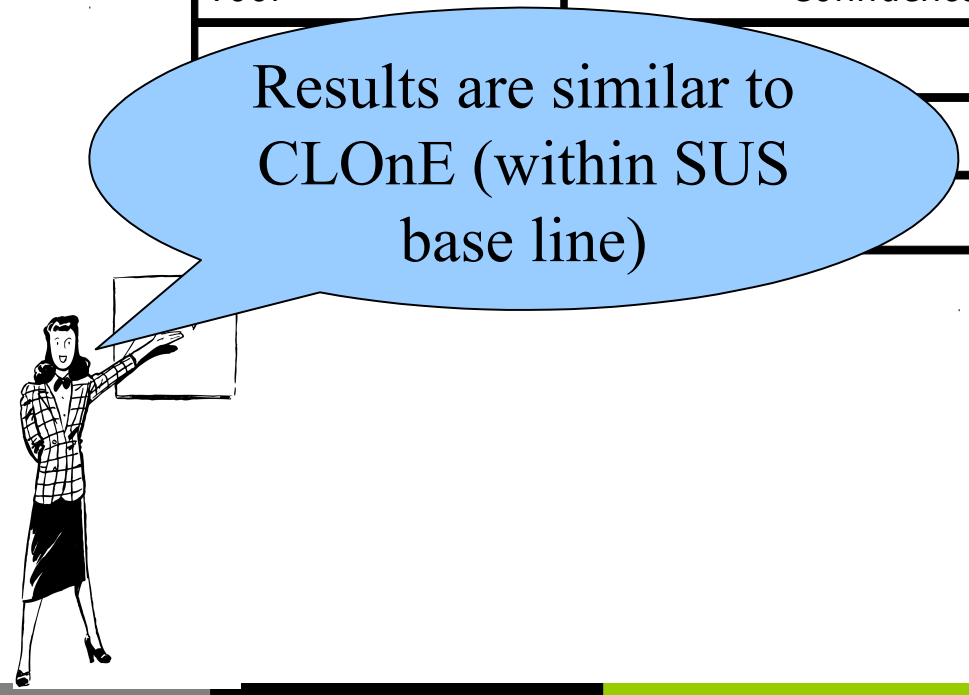
Backup Slides



Evaluation

95% confidence intervals of SUS scores (SUS baseline is 65 to 70%)

Tool	Confidence Intervals	
	Task list B	Combined
	29- 51	32- 49
	69- 84	68- 79



science foundation ireland
fondúireacht eolaíochta éireann
NUI Galway
OÉ Gaillimh
Enabling networked knowledge.

Measure	Measure	Pearson's	Spearman's	Correlation
Pre-test	ROA time	-0.41	-0.21	weak -
Pre-test	Protege time	-0.28	-0.35	none
Pre-test	ROA SUS	-0.02	-0.00	none
Pre-test	Protege SUS	-0.32	-0.29	weak-
ROA time	Protege time	0.53	0.58	+
ROA time	ROA SUS	-0.65	-0.52	-
Protege time	Protege SUS	0.53	0.56	+
ROA time	Protege SUS	-0.14	-0.10	none
Protege time	ROA SUS	-0.02	-0.09	none
ROA SUS	Protege SUS	0.04	-0.01	none
ROA SUS	R/P Preference	0.58	0.56	+
Protege SUS	R/P Preference	-0.01	0.10	none

science foundation ireland fondúireacht eolaíochta éireann
NUI Galway OÉ Gaillimh
Enabling networked knowledge.

Shallow NLG – Pros and Cons

Advantages

- easy to use for non-experts (non NLG)
- easy for domain experts to understand.
- easy and fast to implement

Disadvantages

- difficult to maintain
- Extending is expensive
- output is OK
- cannot handle a domain shift

Deep NLG – Pros and Cons

Advantages
- Maintainable
- Improved Text quality
- Multilinguality
- Standard conformance

Disadvantages
- Need experts
- Overgeneration
- Difficulty scaling