

JISC

KING'S  
*College*  
LONDON  
University of London

CeRch  
Centre for e-Research



The  
University  
Of  
Sheffield.



# TEXTvire: implementing an institutional TextGrid

Mark Hedges & José Miguel Vieira  
Centre for e-Research & Dept of Digital Humanities,  
King's College London, UK

# A message for our funders

- Funded by JISC (Joint Information Systems Committee)
- Part of their VRE (Virtual Research Environment) programme
- Runs until September 2011

JISC



The  
University  
Of  
Sheffield.





# Background

# Starting Point



Collaborative environment for textual scholarship, using German national infrastructure



Research practices  
(to be) supported by  
institutional infrastructure

# Institutional Background

- Research context
  - Department of Digital Humanities
  - Textual scholarship, online editions, TEI XML
  - Collaborations between scholars and technology experts (and other institutions)
- Infrastructural context:
  - Institutional repositories
  - Preservation environment
  - Research management (REF)

# Integration and embedding

- These contexts are quite separate
- Integrate with institutional infrastructure
- Embed within the day-to-day work practices of the researchers



**TEXTv**re





# Users

# Researcher practice

- Scholars do their own thing, so ...
- Explore current and potential research practices and use of tools
  - Semi-structured interviews
  - Tool walk-through (what could they do?)
- Integrate tools to support and enhance these practices (current and potential)

# Research Partners

- Early English Laws
  - Heterogeneous sources
  - Distributed researchers
- Inscriptions of Roman Tripolitania
  - Established editing tradition
  - Material sources
- Gascon Rolls
  - Records editing tradition
  - Entity analysis





# Tools

# Tool Integration

- oXygen XML Editor
- Search (Solr)
- Entity Management
- Named Entity Recognition

# Search (Solr)

User can choose:

- Which documents to index
- How the document contents are indexed
  - Indexing XSLT
  - Solr schema

Search:

- Apache Lucene query syntax
  - name:john
  - name:john AND place:london

# Entity Management - 1

## Entity Authority Tool Set: EATS

- <http://code.google.com/p/eats/>

Perform look-up of selected text in EATS, for identifying entities

Annotate text with entity details

`<name key="entity-003061" type="place">Portchester</name>`

`<name key="entity-000781" type="person">John de Stonor</name>`

## Entity Management - 2

Create new entities directly from TEI markup

Each entity assigned an identifier that can be mapped to a URI

Indices of entities in documents

- Link to occurrence
- Link to entity page in EATS

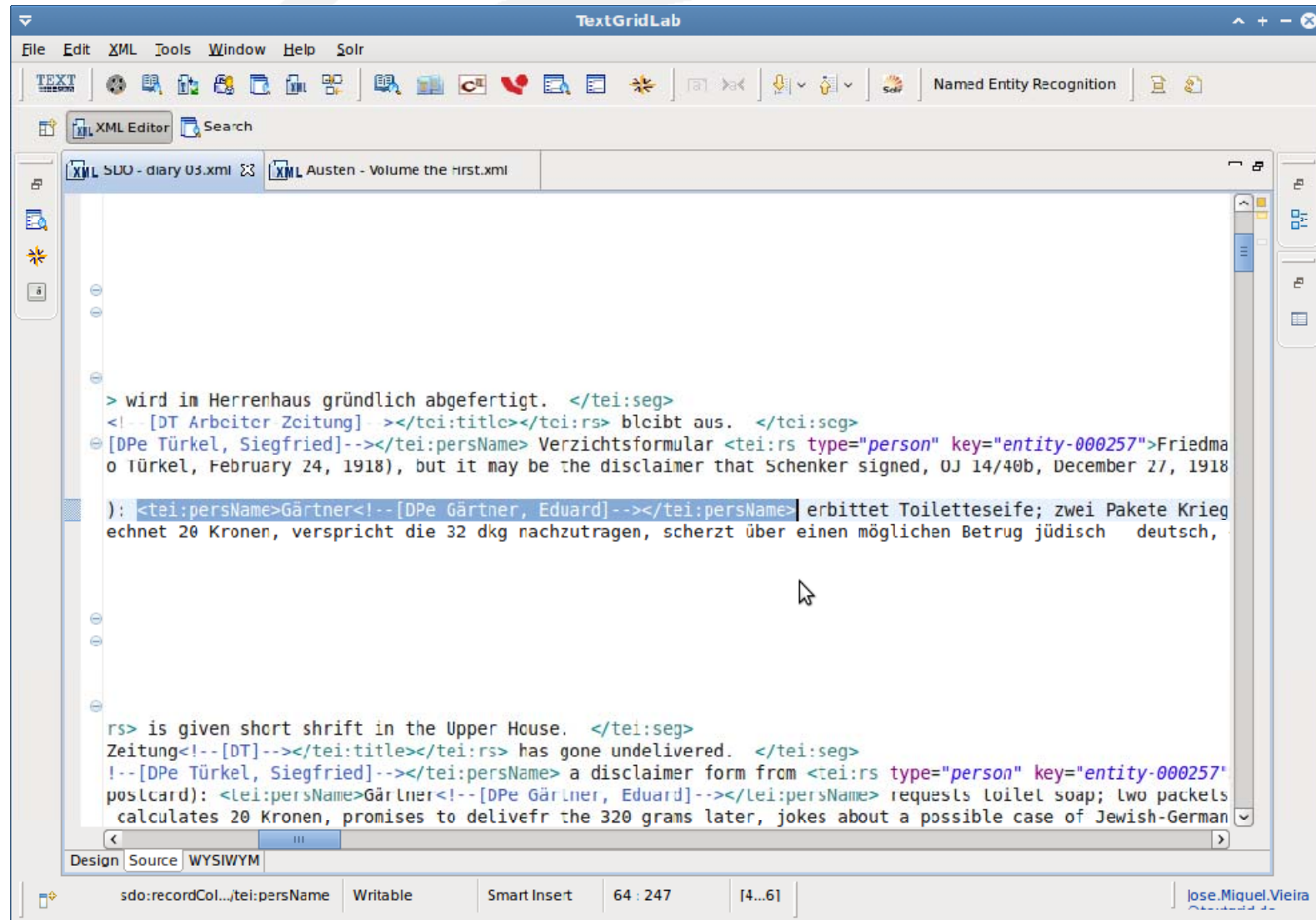
Export from EATS:

- topic maps
- XML ("EATS XML")

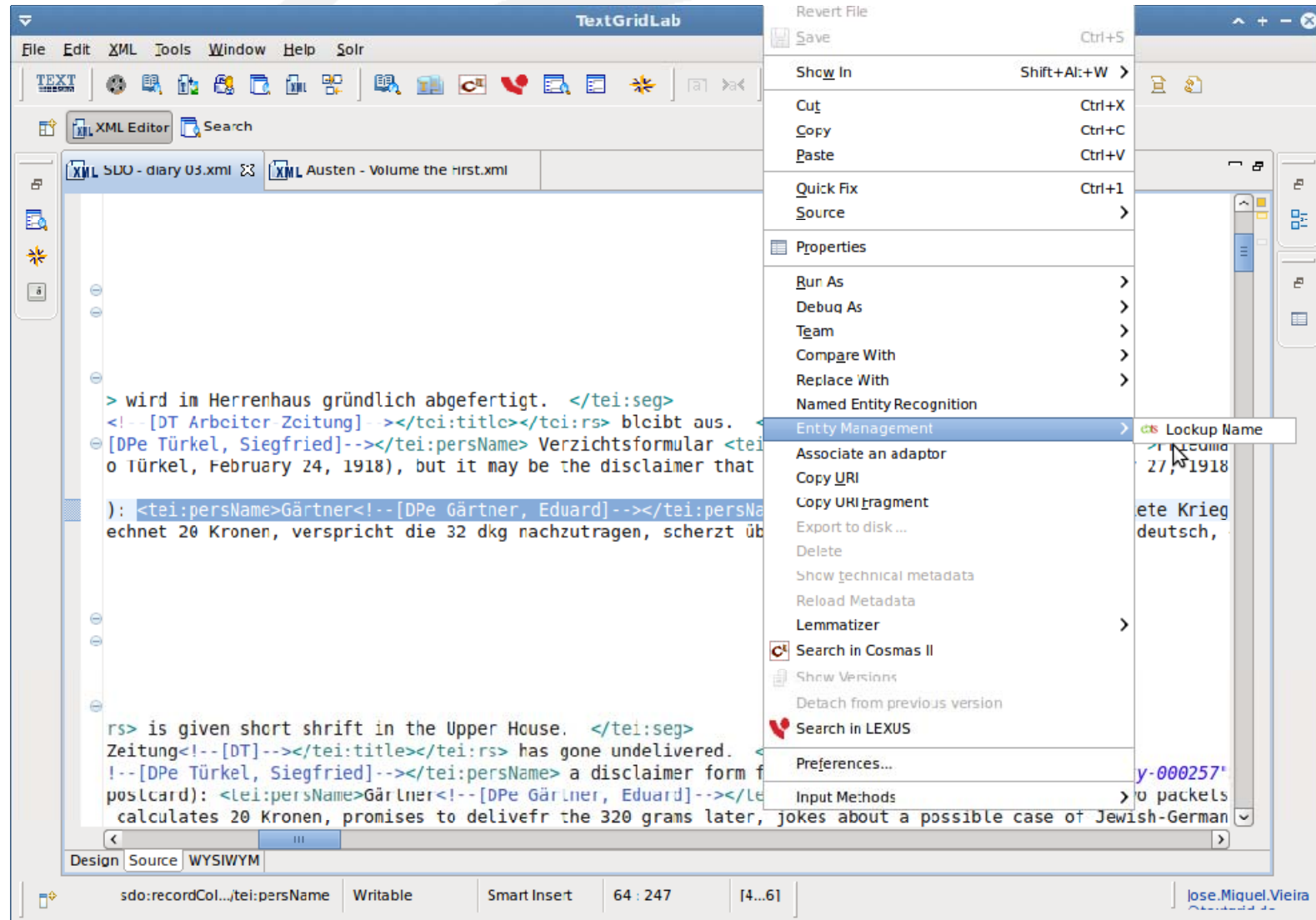
# Named Entity Recognition

- Integration with GATE services
- Automatically add entity markup to documents
- Automatically add entities to EATS
- Currently WIP

# EATS Screenshots - 1



# EATS Screenshots - 2



# EATS Screenshots - 3

Entity details table:

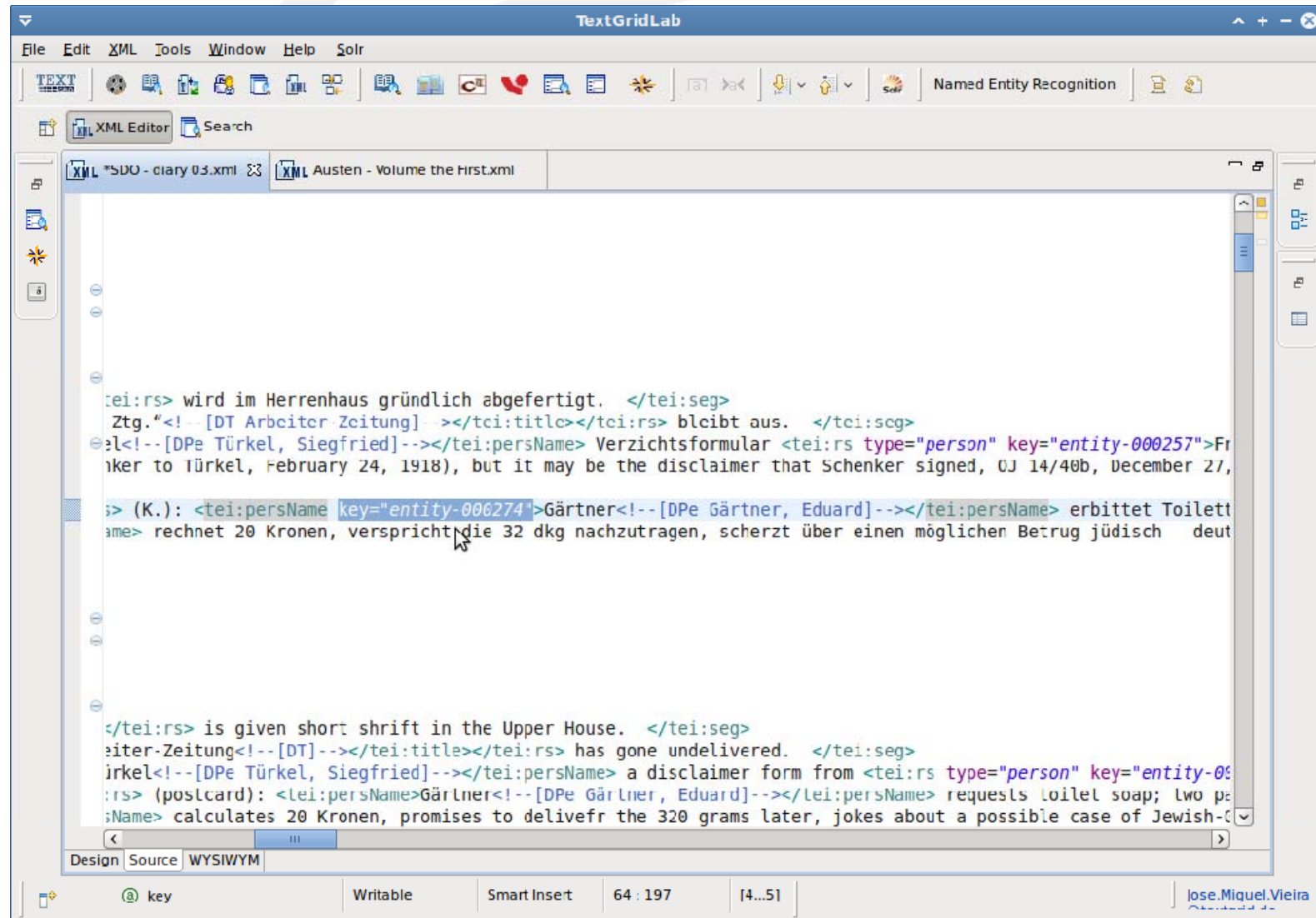
Entity details	Entity type	Family	Given	Terms o...
Eduard Gärtner	person	Gärtner	Eduard	
Emil Gärtner	person	Gärtner	Emil	
Ernst Gärtner	person			
Correspondence between ...	corresp...			
Hans Gärtner	person	Gärtner	Hans	
Poldi Gärtner	person	Gärtner	Poldi	

Authority Records

[entity-000274](#)

Found 6 entities

# EATS Screenshots - 4



# EATS Screenshots - 5

EATS: Entity display: Eduard Gärtner



[Logout](#)

## Eduard Gärtner

[Edit](#)

[Advanced Search](#)

### Authority records

Every piece of information about an entity is associated with an authority record.

Schenker Documents Online record [entity-000274](#)

### Types

person [SDO record entity-000274]

### Names

Eduard Gärtner [SDO record entity-000274]  
regular, German, Latin

No derivable name exists [SDO record entity-000274]  
list id, English, Latin



# Institutional

# Institutional infrastructure



Researcher work area,  
UI, tools, services

---



TextGrid server  
software (adapted)

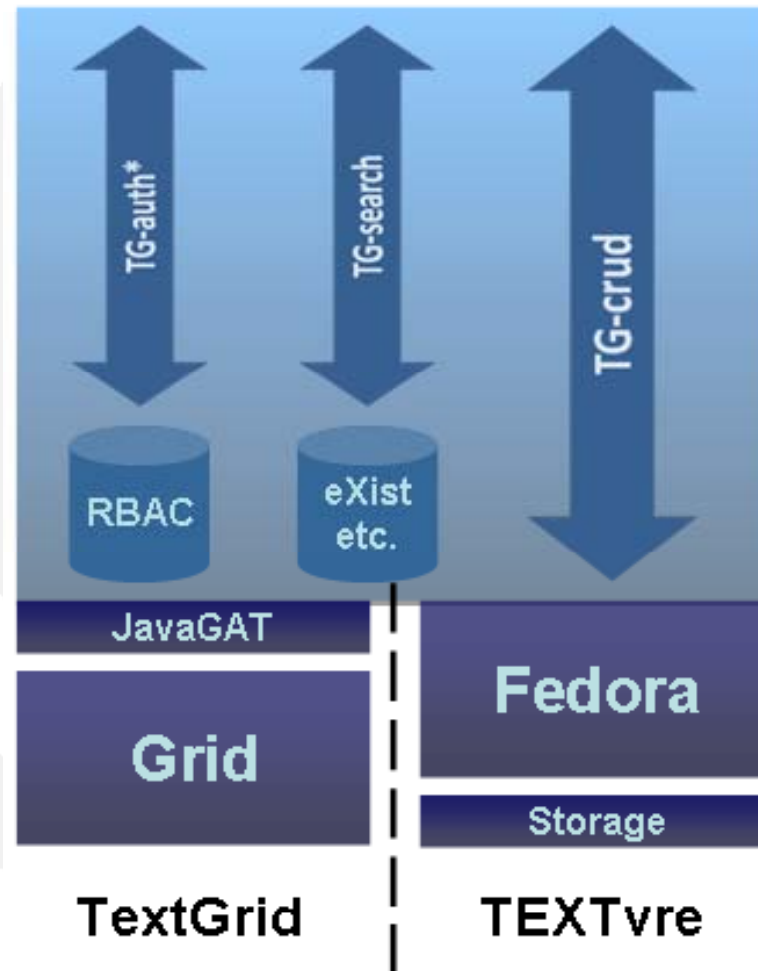
---

Institutional repositories

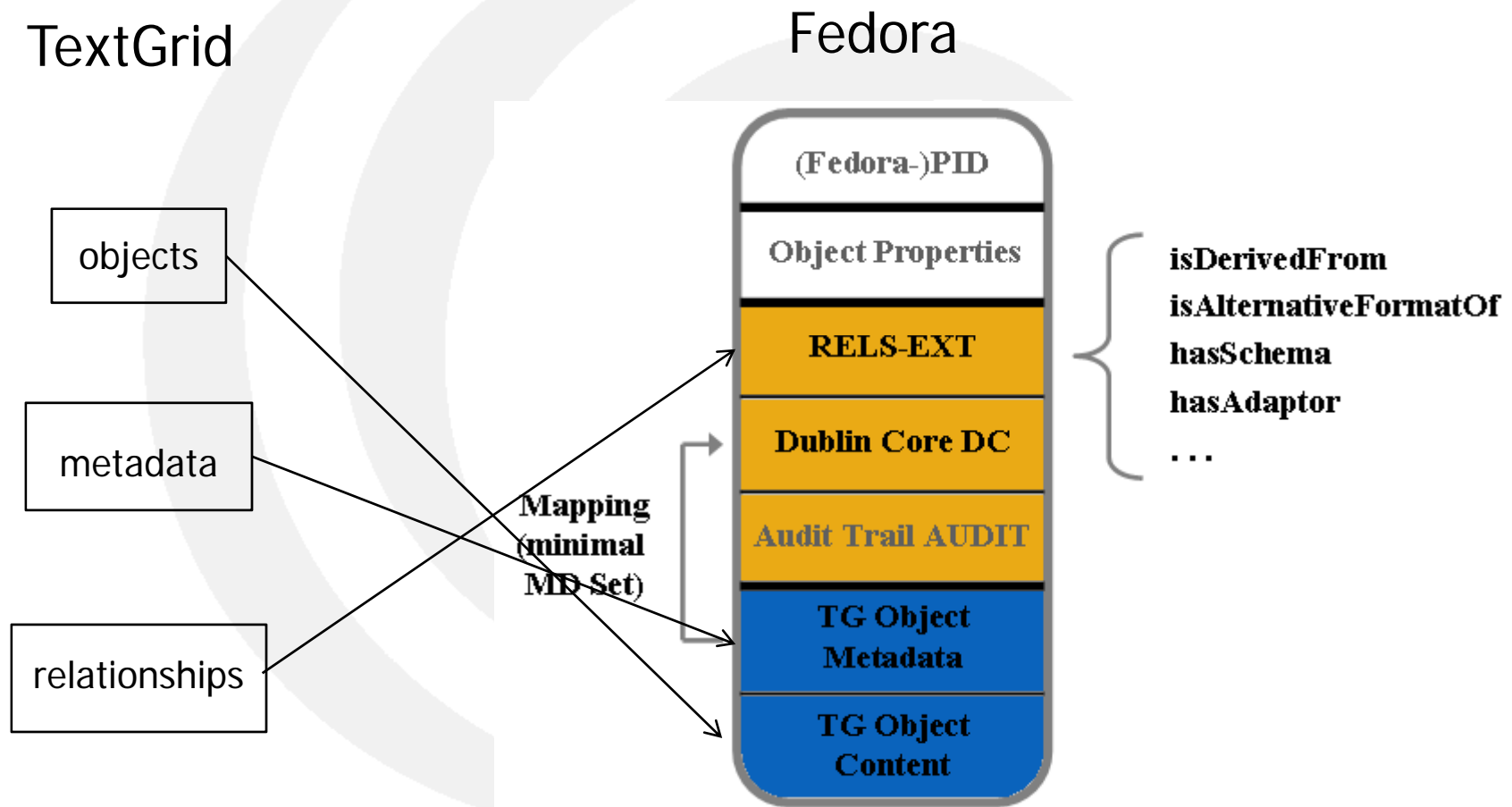


Data curation and preservation

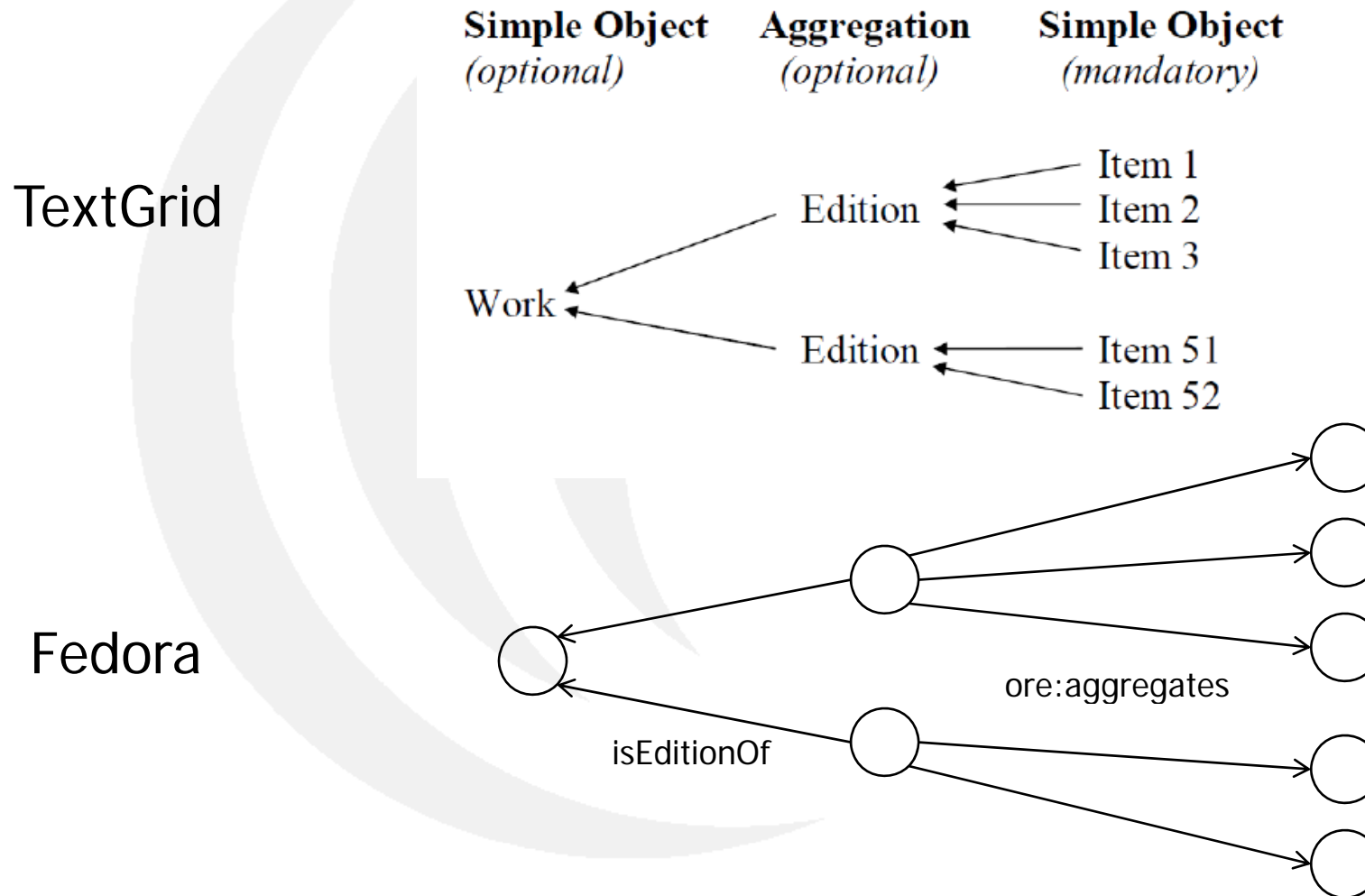
# Fedora Integration - 1



# Fedora Integration - 2



# Fedora Integration - 3





Future

# New Horizons ...

- GATE
  - Further integration
- OCRpodium
  - OCR (Optical Character Recognition)
  - Text extraction from digitised images
- Additional user projects:
  - SAWS (HERA-funded project)
  - Text-Text Link Editor
- Fedora integration:
  - Modularisation of TG-CRUD.

# GATE

The screenshot shows the GATE software interface. On the left, a project tree lists various components: Applications (ANNIE, Corpus Pipeline\_00027), Language Resources (GATE Corpus\_0002A, 8511857.stm\_00021), Processing Resources (StanfordParser\_0001D, ANNIE OrthoMatcher, ANNIE NE Transducer, ANNIE POS Tagger, ANNIE Sentence Splitter, ANNIE Gazetteer, ANNIE English Tokeniser, Document Reset PR), and Datastores.

The main text area displays a document with various annotations. A context menu is open over the text, showing options like 'Default', 'Original markups', 'Lookup', 'Location', 'Token', and 'Sentence'. The 'Location' option is highlighted, and a sub-menu is visible with 'Delete' selected.

On the right, a 'Default annotations' panel lists various annotation types with checkboxes indicating which are active:

- Date
- FirstPerson
- Foo
- Location
- Lookup
- Organization
- Person
- Sentence
- SpaceToken
- Split
- Temp

The text in the main area includes: "was an evidence who swore, he saw him burn them halt a Yard long) and burn them at Singleton's House; which not b Cause for an Indictment, they were acquitted. Andrew James, (a little Boy) of the Parish of St. Dunstun in the West, wa stealing a Silk Handkerchief, value 2 s. from the P of George Mac, on the 8th instant. It was prov'd that the kerchief taken upon him; whereupon he was found Guilty to the Value of 10 d. Mary was indicted for Assaulting ) with infection to on the 2nd of November last. It appear'd by

# OCRopodium

The screenshot shows the OCRopodium web interface. At the top, there is a navigation bar with 'Media', 'Job List', 'Servers', and 'Quick OCR'. Below this, there is a form for submitting a file. The 'Image File' field is empty, and the 'Ocropus version' is set to '0.0.0'. There are two 'Add parameter' fields: one with 'debug' and 'info,transcript', and another with 'debug\_seg' and '/home/michael/alice1\_se'. A 'Submit' button is visible. Below the form, the 'Ocropus returned:' section displays the OCR output for a document snippet, which is a paragraph from 'Alice's Adventures in Wonderland'.

**Ocropus returned:**

1 Down the Rabbit-hole Alice was beginning to get very tired of sitting by her sister on the bank: and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, & when suddenly a White Rabbit with pink eyes ran close by her, There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (When she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural): but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet; for it flashed across her mind that she had never before seen a rabbit with either a waistcoat pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge. In another moment down went Alice after it: never once considering how in the world she was to get out again. The rabbit-hole went straight on like a tunnel, for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well. Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her and to wonder what was going to happen next. First, she tried to look down and make out what she was coming to; but it was too dark to see anything; then she looked at the sides of the well, and noticed that they were filled with cupboards and book-shelves; here and there she saw maps and pictures hung upon pegs. She took down a jar from one of the shelves as she passed; it was labelled 'ORANGE MARMALADE', but to her great disappointment it was empty: she did not like to drop the jar for fear of killing somebody, so managed to put it into one of the cupboards as she fell past it. 'Well!' thought Alice to herself, 'after such a fall as this, I shall think nothing of tumbling down stairs! How brave they'll all think me at home! Why! I wouldn't say anything about it! even if I fell off the top of the house!'

Image File: alice\_1.png  
Ocropus version: 0.0.0  
Time taken: 18 secs  
Extra params: [{"name": "debug", "value": "info,transcript"}, {"name": "debug\_seg", "value": "/home/michael/alice1\_segment\_debug.png"}]

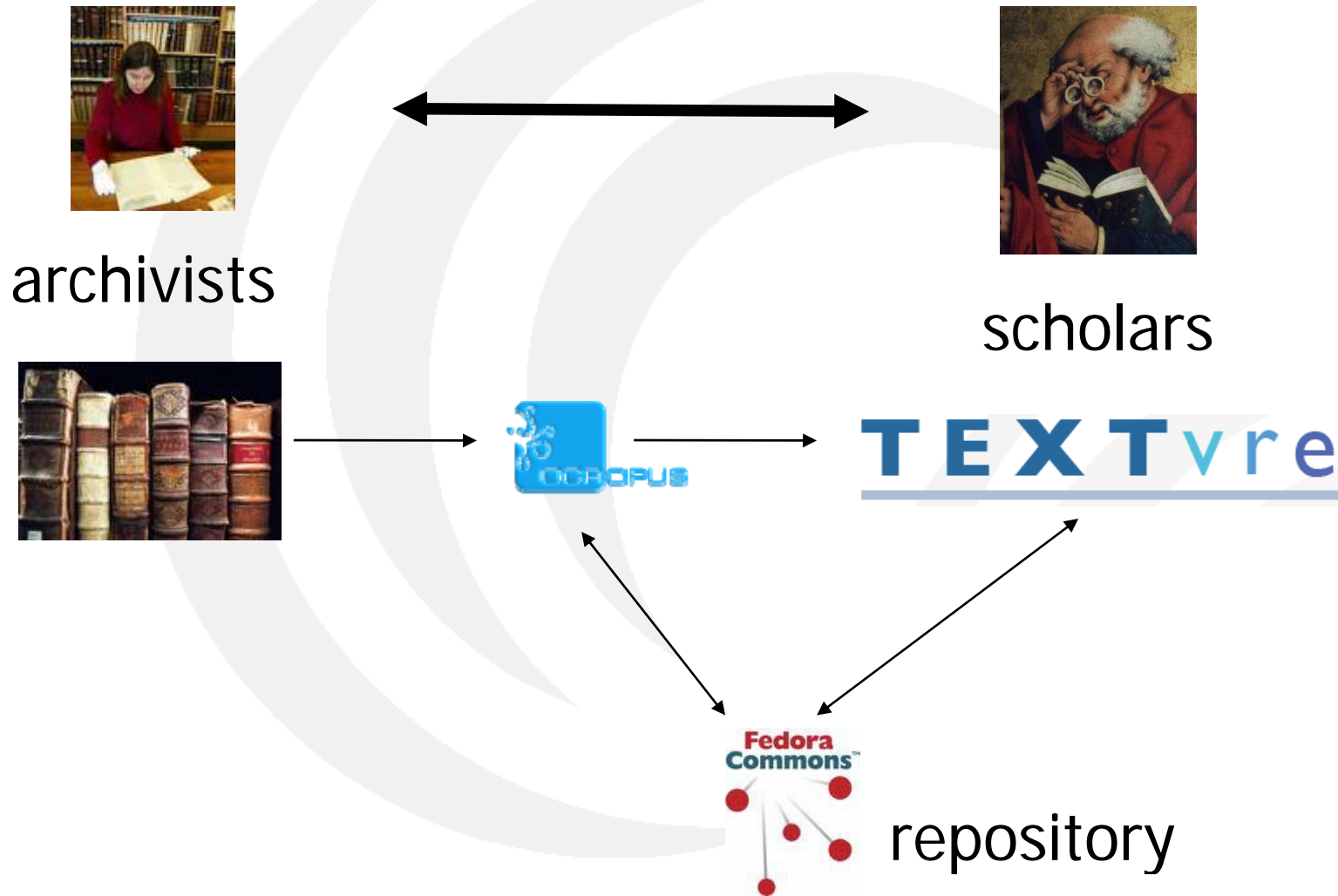
The screenshot shows the OCRopodium web interface with a media pool listing. The listing includes columns for 'Directory', 'Type', 'Files', and 'Submit Batch'. A preview window is open, showing a document page with text and a table.

Directory	Type	Files	Submit Batch
00400	TIF	3	Submit Batch
00200	TIF	10	Submit Batch
00300	TIF	76	Submit Batch
00100	TIF	4	Submit Batch
ocropus test	PNG	186	Submit Batch
00300	TIF	140	Submit Batch

Media: 8 The online historical population reports 1801-1937

Preview | Next | Submit Page

# More institutional integration



## To summarise

- Dispersed scholars working on diverse (digital) humanities projects
- Develop a VRE that is embedded in the day-to-day research activities of scholars
- Integrate VRE with institutional infrastructure(s): repositories, preservation, archives/library
- Provide integrated framework for dealing with (text-based) historical and archival material

# Contacts

<http://textvire.cerch.kcl.ac.uk/>

[mark.hedges@kcl.ac.uk](mailto:mark.hedges@kcl.ac.uk)