CLARIAH WP3: Virtual Research Environment & provenance

Menzo Windhouwer
menzo.windhouwer@di.huc.knaw.nl

Bas Leenknecht
QiQing Ding
Daan Broeder
Outline

• Overview of the CLARIAH WP3 VRE
• The VRE and provenance
  • Functionality for users
  • Support by the VRE
  • Support by the infrastructure
• Conclusions & future work
Why a CLARIAH WP3 VRE?

- Overcoming the current fragmentation of linguistic services
- Connecting to CLARIN/CLARIAH practices and infrastructure components e.g. CMDI, Switchboard, ...
- Gain experience supporting research data workflows and include (automatic) facilities for open and reproducible science e.g. data provenance & easy archiving and publication
- Integration of remote services, including a service deployment framework improving the TTNWW experience
- Offering UI suitable for the well known 80% of non-tech users, but not excluding the possibilities of cl and console for the 20%
VRE at a glance

An integrative platform for the humanist

1. **Upload your resources** to your workspace
   - Automatic data type detection, (semi)automatic component metadata creation, automatic parsing and indexing, ...

2. **Invoke a matching service** to process the resources
   - Using a plugin architecture many service invocation methods can be supported, e.g., GATE Cloud, docker, WebLicht and **CLAM**
   - Stanbol, Gretel, Paqu, **Ucto**, Frog, ...

3. **Results come available in your workspace** to be further processed or downloaded

4. Archive relevant results into a repository, e.g., the Meertens TLA-FLAT or EUDAT’s B2SHARE
   - **Provenance is tracked**, i.e., its known which and how services were used to process the data
VRE at a glance

User Interface

Internal switchboard

Deployment

Registers & Provenance logs

GATE

CLAM

docker

(Ucto)

(Web) Services

Repository

Provenance workshop – 3 September 2018 - slide 4
VRE and provenance – functionality for users

• Maintain and inspect a journal of all actions that happened to a resource
  • From primary source material, e.g., a speech recording or a manuscript
  • To final secondary data, e.g., an ELAN or FoLiA resource that has been created by a mix of automatic generation and manual clean-up and validation
  • And anything in-between
• Serialize this to a standard representation, e.g., W3C PROV, to be included with a deposit of the data as part of the scientific (data) publication process
  • Supporting the goal of reproducible science
### Provenance of NOS.xml a FoLiA resource @ 2018-09-02 14:09:10

<table>
<thead>
<tr>
<th>When?</th>
<th>What?</th>
<th>How?</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018-09-01 20:15:02</td>
<td>NOS.xml a FoLiA resource</td>
<td>✍️ edited by menwin in FLAT: fixed some PoS tags</td>
</tr>
<tr>
<td>20:10:05</td>
<td>NOS.xml a FoLiA resource</td>
<td>✔️ processed with FROG</td>
</tr>
<tr>
<td>2018-08-31 13:40:55</td>
<td>NOS.txt a Plain text resource</td>
<td>✔️ processed with UCTO{language=Dutch}</td>
</tr>
</tbody>
</table>
Support by the VRE

- Know which service was run to produce an output resource from an input resource, preferably with
  - Parameters
    - Available as the VRE asks them from the user or configures them
  - version info
    - Available, as far as (remote) services don’t get out of sync and don’t provide provenance themselves

- The CLARIAH WP3 VRE stores this in a combination of its object and service registries and a permanent event log, i.e., Apache Kafka

- Know when resources are updated outside the control of the VRE
  - on the local machine of the user

- The CLARIAH WP3 VRE monitors a local directory which kept in sync with its server, i.e., via OwnCloud
  - ask the user to specify the local changes (in an unobtrusive way)
Support by the infrastructure

• Services should
  • provide provenance information
    • unique version of a single service or fingerprint of a pipeline
  • make specific versions persistently available
    🟠 by being still deployed
    🟡 as a docker image in a trusted repository
    🟢 as source code in a trusted repository

• Repositories should welcome provenance next to resources

Provenance workshop – 3 September 2018 - slide 8
Conclusion & future works

• The CLARIAH WP3 VRE works towards
  • supporting users in preparing and analysing their resources and keeping track of what has been done
  • supporting the goal of reproducible science
    • which needs broad support to be actually achieved

• VRE status
  • Basic functionality available incl. all registries and logs to support provenance
  • Adding support for viewers and editors
  • Start working on services and UI to access and maintain the collected provenance
Lessons already learned

• How much (intermediate) versions to keep?
  • delta/diffs

• Keep (external) searches (history)
  • Some support in the VCR, Nederlab

• Editable search results, i.e., indicate the results you keep

• How deeply document manual steps?
  • VRE being unobtrusive?

• Handle (keep) error logs
  • Indicate what is (manually) fixes, or is actually correct

• Fingerprint software
  • Wrapup all dependencies in containers
  • In case of source code how deep do we go?
VRE – guided tour

Welcome to Clariah’s VRE
VRE – guided tour

A CLARIAH WP3 Virtual Research Environment

<table>
<thead>
<tr>
<th>#</th>
<th>File</th>
<th>Format</th>
<th>Mimtype</th>
<th>Created</th>
<th>User</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>admin</td>
<td>directory</td>
<td>inode/directory</td>
<td>2018-08-30 12:14:32.857541+00</td>
<td>admin</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>admin/files/output-2018-08-30_12-14-23-271/result.txt</td>
<td>Plain text</td>
<td>text/plain</td>
<td>2018-08-30 12:14:35.112062+00</td>
<td>admin</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>admin/files/output-2018-08-30_12-14-26-415/result.txt</td>
<td>Plain text</td>
<td>text/plain</td>
<td>2018-08-30 12:14:35.623329+00</td>
<td>admin</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>admin/files/test-94b781b6-0927-4ff1-9256-4f47c8347e0c0-20180830_121550.txt</td>
<td>Plain text</td>
<td>text/plain</td>
<td>2018-08-30 12:15:54.145499+00</td>
<td>admin</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>admin/files/test-61ca148e-1b18-4f3b-bb52-0c4377e8b232-20180830_121628.txt</td>
<td>Plain text</td>
<td>text/plain</td>
<td>2018-08-30 12:16:29.908931+00</td>
<td>admin</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>admin/files/test-31533ac4-d4c2-463f-b6ad-3ca24faef29a-20180830_121629.txt</td>
<td>Plain text</td>
<td>text/plain</td>
<td>2018-08-30 12:16:30.449498+00</td>
<td>admin</td>
<td></td>
</tr>
</tbody>
</table>
VRE – guided tour
VRE – guided tour
VRE – guided tour

A CLARIAH WP3 Virtual Research Environment

[Image of VRE interface]

1. Select service

Process `result.txt` with one of the services below.

<table>
<thead>
<tr>
<th>#</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TEST</td>
</tr>
</tbody>
</table>

[Image of select service interface]
VRE – guided tour

A CLARIAH WP3 Virtual Research Environment

1. Select service

Process `result.txt` with one of the services below.

<table>
<thead>
<tr>
<th>#</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TEST</td>
</tr>
</tbody>
</table>

Provenance workshop – 3 September 2018 - slide 16
VRE – guided tour

2. Configure service

Test input file

Selected file: /admin/files/output-2018-08-30_12-14-23-271/result.txt

Author

bial
VRE – guided tour

VRE
A CLARIAH WP3 Virtual Research Environment

Home Files

<< Select file Select service Configure service Deploy >>

Deployment of service [TEST] has been requested.

Bekijk status

<< Select file Select service Configure service Deploy >>
VRE – guided tour