

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK A DEVELOPER'S AND USER'S PERSPECTIVE

Sally Chambers¹, Joke Daems¹, Greta Franzini², Marco Büchler², Susan Aasman³

¹Ghent Centre for Digital Humanities, Ghent University

²Institute of Computer Science, University of Göttingen

³Groningen Centre for Digital Humanities, University of Groningen

OVERVIEW

- U4 network and DH4U4
- Towards a tool and data criticism framework
 - Tool evaluation criteria
 - Data evaluation criteria
 - Combined framework
- Validation
- Next steps

U4 NETWORK AND DH4U4

- Strategic partnership between Ghent University (BE), University of Göttingen (DE), University of Groningen (NL) and Uppsala University (SE)
- Platform for collaboration between the four universities
- **Digital Humanities for U4 (DH4U4)**: taskforce within the U4 Humanities Cluster established in November 2015
- Stimulate exchange of Digital Humanities knowledge and expertise between the U4 universities

DH4U4 ACTIVITIES

- **Collaborative project proposals:** Computational Social Sciences and Humanities
- **Staff exchanges:** Marco Büchler's research visit to Ghent (Nov 2016) and to Groningen (Feb 2017), Joke Daems participation in DATeCH conference in Göttingen (June 2017), Melina Jander's research fellowship in Ghent (Sep-Nov 2017), Jules de Doncker research fellowship in Göttingen (awaiting result)
- **Joint Master's supervision:** Groningen and Göttingen: Peter Sprenger
- **Co-publications:** joint presentation at DH Benelux 2017
- **Next steps:** DH4U4 Doctoral Schools programme (mobility of PhD students for DH doctoral training)

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

- **Sally Chambers:** *Digital Humanities Research Coordinator*. Expertise: metadata and research data management.
- **Joke Daems:** *Translation Studies*. Research: Digital Text Analysis, Translation Studies.
- **Susan Aasman:** *Media Historian*. Research: Media History, Digital History, Everyday Digital Practices.
- **Marco Büchler:** *Computer Scientist*. Research: Natural Language Processing, Big (Humanities) Data, Text Reuse.
- **Greta Franzini:** *Classicist*. Research: Digital Classics, Digital Editing, Natural Language Processing.

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

- **Digital Collections as Data**, e.g. *Delpher Newspapers*' collection from the National Library of the Netherlands, or AV Collections Netherlands Institute for Sound and Vision
- **Digital Tools**, e.g. DiRT Digital Research Tools directory, DARIAH, CLARIN, CLARIAH ...
- **Need for a framework that:** a) takes into account **both tool *and* data** used, b) facilitates **better communication between developers and users**
- **DH4U4:** framework to **facilitate DH peer-review** between our universities

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

BUILDS ON:

- ‘Tool Criticism for Digital Humanities’ workshop (Traub and Ossenbruggen, 2015)
- ‘Source criticism’ and ‘information evaluation’ frameworks (Hjorland, Birger 2012)
- Analogous software studies (Jackson et al., 2011)
- EVALITA (Evaluation of NLP and Speech Tools for Italian) campaigns
- RIDE Digital Text Collections evaluation guidelines

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

PROPOSED TOOL EVALUATION CRITERIA - 1

1. Usability

- a. User Experience (UX)
- b. Graphical User Interface (GUI)

2. Documentation

- a. Provenance (authors / organisations behind the tools)
- b. “How to instructions”
- c. Algorithms or methods implemented
- d. Limitations
- e. Target audience/research
- f. Availability of tutorials to train users to proficiently work with the tool
- g. Access and citation
- h. Rights

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

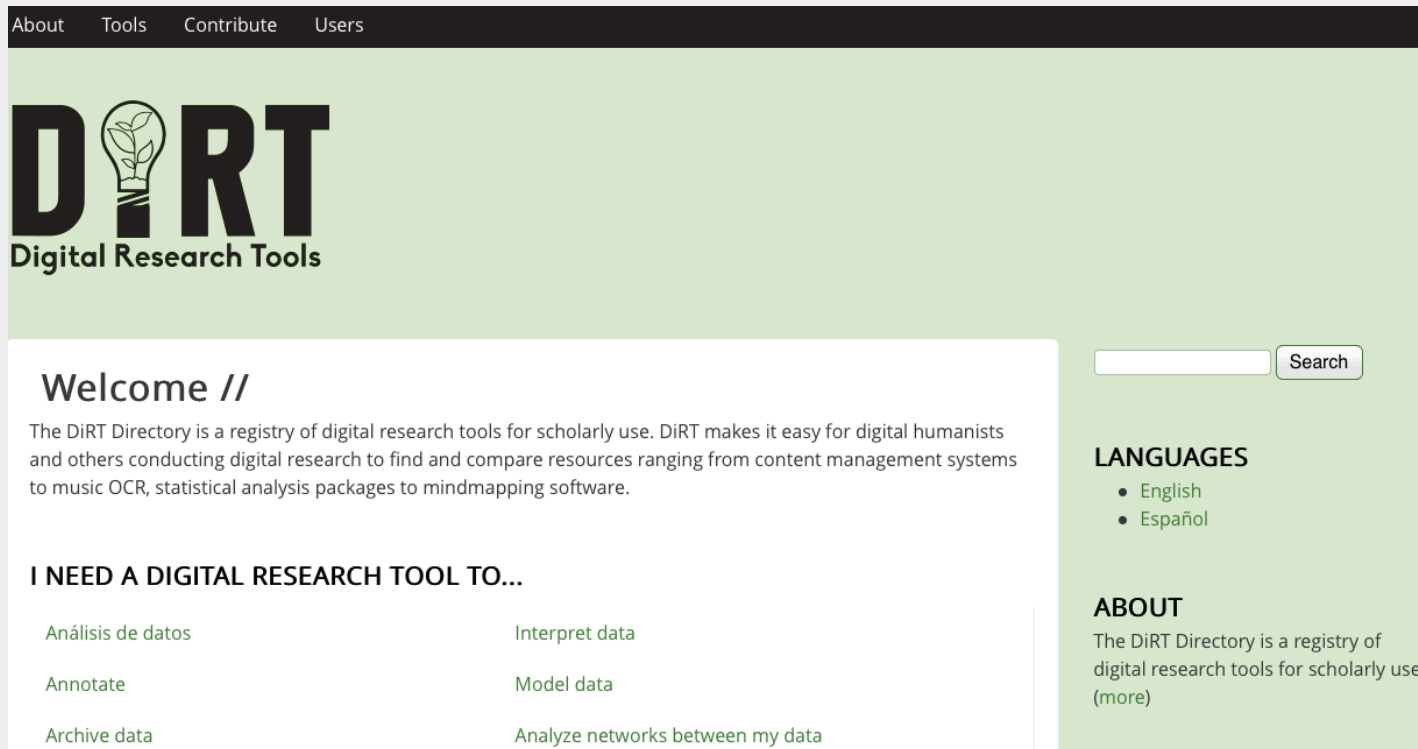
PROPOSED TOOL EVALUATION CRITERIA - 2

3. Sustainability and Maintenance

- a. Development responses to user feedback
- b. Preventing 'tool rot' (i.e. if you have been using a tool and then the development stops and you are left with bugs and eventually an unusable tool)

4. Flexibility/Extent of Applicability

INCORPORATE INTO DiRT DIRECTORY?



The screenshot shows the DiRT Directory website. At the top is a dark navigation bar with links: About, Tools, Contribute, Users. Below this is a light green header area containing the DiRT logo (a lightbulb with a plant inside) and the text "Digital Research Tools". The main content area is white and features a "Welcome //" section with a paragraph describing the directory. To the right is a search bar with a "Search" button. Below the search bar is a "LANGUAGES" section with links for English and Español. At the bottom right is an "ABOUT" section with a paragraph and a "(more)" link. On the left side of the main content area, under the heading "I NEED A DIGITAL RESEARCH TOOL TO...", there is a two-column list of tool categories: "Análisis de datos", "Interpret data", "Annotate", "Model data", "Archive data", and "Analyze networks between my data".

About Tools Contribute Users

DiRT

Digital Research Tools

Welcome //

The DiRT Directory is a registry of digital research tools for scholarly use. DiRT makes it easy for digital humanists and others conducting digital research to find and compare resources ranging from content management systems to music OCR, statistical analysis packages to mindmapping software.

I NEED A DIGITAL RESEARCH TOOL TO...

Análisis de datos	Interpret data
Annotate	Model data
Archive data	Analyze networks between my data

Search

LANGUAGES

- English
- Español

ABOUT

The DiRT Directory is a registry of digital research tools for scholarly use. (more)

<http://dirtdirectory.org>

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

PROPOSED DATA EVALUATION CRITERIA - 1

1. (Re-)Usability

- a. Format(s)

2. Documentation

- a. Provenance (curators / organisations behind the data-sets)
- b. Metadata (e.g. size, source, author, etc.)
- c. Limitations
- d. Access and citation
- e. Rights

3. Sustainability and Maintenance

- a. Development responses to user feedback

TOWARDS A TOOL AND DATA CRITICISM FRAMEWORK

COMBINED FRAMEWORK

TOOLS	DATA
1. Usability	1. (Re-)Usability
2. Documentation	2. Documentation
3. Sustainability & Maintenance	3. Sustainability & Maintenance
4. Flexibility/Extent of Applicability	

VALIDATING THE PROPOSED FRAMEWORK

TRACER: AUTOMATIC TEXT REUSE DETECTION

- Advance research in automatic text reuse detection in historical texts (small and large corpora)
- Transparent detection process
- Tune it to the needs of humanists and literary scholars with little to no knowledge/experience in NLP
- Integration with existing linguistic resources for historical languages (e.g. TreeTagger, Stanford CoreNLP)
- Turn it into a web-service



<http://www.etrp.eu/research/tracer/>

CRITERION 1b) GRAPHICAL USER INTERFACE

User's perspective

- Con: TRACER doesn't come with a Graphical User Interface (GUI)
- Pro: Output visualisations can be generated, such as a Dotplot view, a Variant graph, etc.

```

Writing meta information ... DONE!

END OF PROCESS LEVEL 3 (SELECTION)

START TO PROCESS LEVEL 4 (LINKING)

Using de.gcdh.tracer.linking.IntraCorpusLinkingImpl implementation for linking.
OUTPUT file is outfile-data/corpora/Bible/TRACER_DATA/01-02-WLP-lem_true_syn_true_ssim_false_redwo_false-ngram_5-LLR_true_toLC_false_rDia_false_w2wl_false-wlt_5/01-02-01-01-
BiGramShinglingTrainingImpl/02-02-01-01-01-LocalMaxFeatureFrequencySelectorImpl_FeatDens=0.8/01-01-01-01-01-KJV-01-01-02-KJV/KJV-KJV.link ...
  Preparing RUID2Feature connector by implementation in de.gcdh.tracer.linking.connector.ram.RUID2FeatureRAMConnectorImpl
  DONE!
  Preparing Feature2RUID connector by implementation in de.gcdh.tracer.linking.connector.ram.Feature2RUIDRAMConnectorImpl
  DONE!
  Linking process started for 28632 fingerprinted re-use units
  28632 re-use units processed. 100% DONE!
  DONE!!

END OF PROCESS LEVEL 4 (LINKING)

START TO PROCESS LEVEL 5 (SCORING)

Using de.gcdh.tracer.scoring.FeatureSelectedSymmetricSelectedFeatureResemblanceSimilarityImpl implementation for scoring.
OUTPUT file is outfile-data/corpora/Bible/TRACER_DATA/01-02-WLP-lem_true_syn_true_ssim_false_redwo_false-ngram_5-LLR_true_toLC_false_rDia_false_w2wl_false-wlt_5/01-02-01-01-
BiGramShinglingTrainingImpl/02-02-01-01-01-LocalMaxFeatureFrequencySelectorImpl_FeatDens=0.8/01-01-01-01-01-KJV-01-01-02-KJV/02-02-01-01-02-SelectedFeatureResemblanceSimi-
larityImpl_Threshold=0.9/KJV-KJV.score ...
  Preparing data from data/corpora/Bible/TRACER_DATA/01-02-WLP-lem_true_syn_true_ssim_false_redwo_false-ngram_5-LLR_true_toLC_false_rDia_false_w2wl_false-wlt_5/01-02-01-01-
1-01-BiGramShinglingTrainingImpl/02-02-01-01-01-LocalMaxFeatureFrequencySelectorImpl_FeatDens=0.8/KJV.sel ...
  Scoring data
  FROM data/corpora/Bible/TRACER_DATA/01-02-WLP-lem_true_syn_true_ssim_false_redwo_false-ngram_5-LLR_true_toLC_false_rDia_false_w2wl_false-wlt_5/01-02-01-01-Bi-
GramShinglingTrainingImpl/02-02-01-01-01-LocalMaxFeatureFrequencySelectorImpl_FeatDens=0.8/01-01-01-01-01-KJV-01-01-02-KJV/KJV-KJV.link ...
  TO data/corpora/Bible/TRACER_DATA/01-02-WLP-lem_true_syn_true_ssim_false_redwo_false-ngram_5-LLR_true_toLC_false_rDia_false_w2wl_false-wlt_5/01-02-01-01-BiGr-
amShinglingTrainingImpl/02-02-01-01-01-LocalMaxFeatureFrequencySelectorImpl_FeatDens=0.8/01-01-01-01-01-KJV-01-01-02-KJV/02-02-01-01-02-SelectedFeatureResemblanceSimilari-
tyImpl_Threshold=0.9/KJV-KJV.score ...
  DONE!!

END OF PROCESS LEVEL 5 (SCORING)

```

CRITERION 1b) GRAPHICAL USER INTERFACE

Developer's perspective

- Text reuse is computationally complex – quadratic-time algorithms $O(n^2)$
- Computational runs can last several hours or even days or weeks
- Computational runs on High Performance Computing (HPC) instead of local laptops
- TRACER tutorials:
 - Teaching not only to use TRACER but also introducing to basic algorithms and...
 - ... to all necessary command line skills, too
 - Result: Participants learn “locally” how to use TRACER but can run it on their home university's HPC cluster, too
- Nevertheless: A TRACER GUI is planned with a micro-grant of Göttingen's campus lab on “Digitisation”

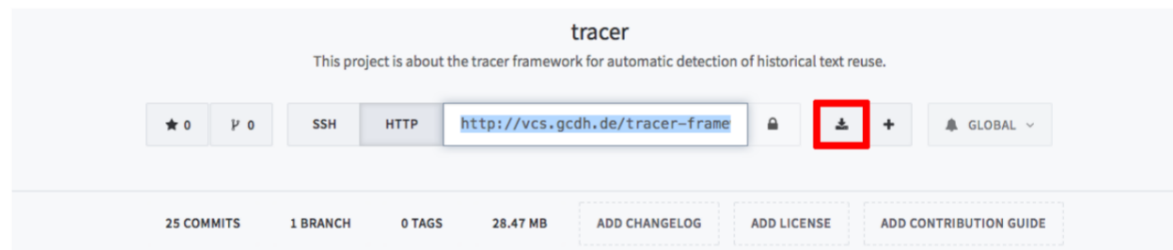
CRITERION 2b) HOW TO “INSTRUCT”?

User’s perspective

- Pro: Evolving user manual of 50+ pages available
- Con: Algorithms are not explained

3.2 Download TRACER with Git

Alternatively, you can clone the most recent releases of TRACER from our [git repository](#)⁴. However, please be aware that the most recent version might be unstable. Once you’ve obtained an account you can download the latest version from [here](#) or by using git.

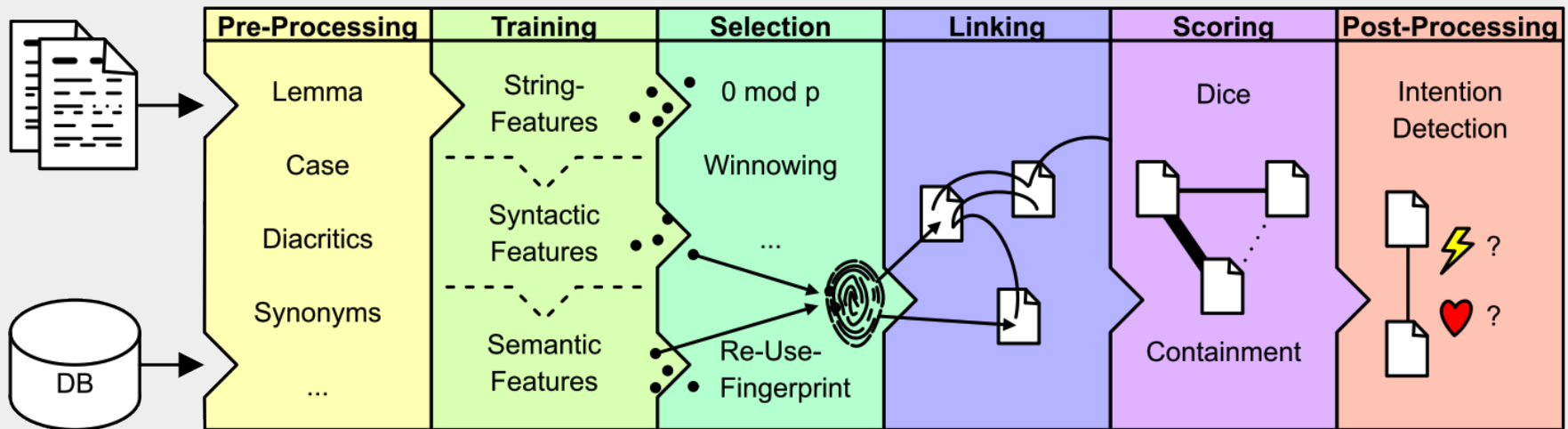


This repository contains the TRACER framework for automatic detection of historical text reuse.

Figure 3.1: TRACER's GitLab repository.

CRITERION 2b) HOW TO “INSTRUCT”?

Developer’s perspective I - Architecture



CRITERION 2b) HOW TO “INSTRUCT”?

Developer’s perspective II – “Debugger”

- TRACER comprises ca. 700 algorithms: If we use ½ page per algorithm, it is no longer a user manual but a 350+ page book
- For this reason: Explaining algorithms by examples
- Nevertheless: handbook is planned

▣ Step 0: Searching

▣ Step 1: Preprocessing

▣ Step 2: Featuring

Please select a training strategy: Bi Gram Shingling Training change

Preprocessed sentence: in the begin god create the heaven and the earth .

Position	Feature
0	in the
1	the begin

Position	Feature
2	begin god
3	god create

Position	Feature
4	create the
5	the heaven

Position	Feature
6	heaven and
7	and the

Position	Feature
8	the earth
9	earth .

next Level

CRITERIA 1a) UX vs. 4) FLEXIBILITY

User's & Developer's perspective

- UX perspective:
 - Easy to use software
 - Intuitive installation & design patterns that do not need “big” explanations
- Flexibility (for different research questions)
 - Need for “algorithmic diversity” and complexity



VALIDATING THE PROPOSED FRAMEWORK

CLARIAH: TRACING FIRST PERSON IN DOCUMENTARY HISTORY IN AV-COLLECTIONS

- Explore the emergence of a genre before it is a well constituted and recognized as such
- Address the challenges of doing historical research in large audiovisual collections by making use of a video annotation tool
- Additionally, this research aims to use contextual sources, like the program guides available in the CoMeRDa tool, to gain more insights
- **And tool and data criticism:** Understand how tools like video annotation and/or the collection explorer work with the *Mediasuite platform* of CLARIAH/Netherlands Institute for Sound and Vision



VALIDATING THE PROPOSED FRAMEWORK



Towards a IIF-based corpus management platform

"I want to perform digital text analysis"



Goals

- Collect data from different possible datastreams
- Generate/extract high-quality textual data
- Search through data to create relevant subcorpus
- Export data for subsequent digital tekst analysis

The envisioned solution

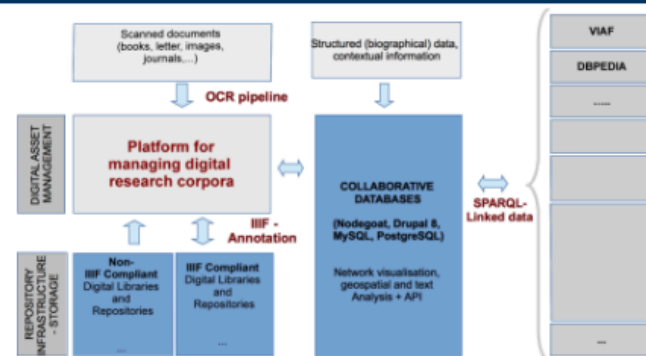
- Import through various datastreams
- OCR pipeline (ingestion + improvement)
- Collaborative addition of metadata and annotations
- Extend the International Image Interoperability Framework to textual data

→ interoperable

→ international standard

→ sharing without exchanging

→ multilingual data



THANK YOU FOR YOUR ATTENTION!

Sally Chambers¹, Joke Daems¹, Greta Franzini², Marco Büchler², Susan Aasman³

¹Ghent Centre for Digital Humanities, Ghent University

²Institute of Computer Science, University of Göttingen

³Groningen Centre for Digital Humanities, University of Groningen



DH Benelux, Utrecht, 3-5 July 2017