OCR-D-Review

Arved Solth (effective Webwork GmbH)

Das Ziel dieses Reviews ist, eine Auswahl von Repositories des OCR-D-Projekts nach mehreren in der Auftragsbeschreibung in Arbeitspaketen formulierten Kriterien zu überprüfen und jeweils eine darauf bezogene Einschätzung abzugeben. Wo gegeben, sollen ermittelte Angaben mit einem vom Auftraggeber zur Verfügung gestellten Pad (https://pad.gwdg.de/XgiEQuPIQeOJQUzJHyeonA#) verglichen werden.

WP 1: Allgemeine Analyse

In diesem Arbeitspaket wurden der aktuelle Zustand sowie die Aktivität der GitHub-Repositories der vier zu berücksichtigenden OCR-D-Komponenten „core“, „ocrd_olena“, „ocrd_tesserocr“ und „orcd_fileformat“ in einem Stand vom 25.5.2024 zum Zeitpunkt dieses Reviews untersucht.

- „core“:
  - 117 offene/379 geschlossene Issues
  - 20/635 Pull Requests
  - Ältester offener Pull Request von 2020
  - 205 Releases nach Semantic Versioning
  - Letzte Code-Änderung: letzte Woche
  - 4140 Commits in master
- „ocrd_olena“:
  - 4/28 Issues
  - 1/63 Pull Requests
  - Ältester offener Pull Request von 2020
  - 22 Releases
  - Letzte Code-Änderung: vor 4 Monaten
  - 263 commits in master
- „ocrd_tesserocr“:
  - 12/76 Issues
  - 4/119 Pull Requests
  - Ältester offener Pull Request von 2020
  - 39 Releases nach Semantic Versioning
  - Letzte Code-Änderung: vor 1 Woche
  - 661 commits
- „ocrd_fileformat“:
  - 4/18 issues
  - 2/28 Pull Requests
  - Ältester offener Pull Request von 03 2024
  - 23 Releases nach Semantic Versioning
  - Letzte Code-Änderung: vor 5 Monaten
  - 112 commits

In einem zweiten Schritt wurden die Entwickler zum Zustand der beobachteten Repositories befragt. Das „core“-Repository erfährt stetige Weiterentwicklung. Es ist geplant, zum Ende des Projektes eine Version 3.0 herauszubringen, die die im Pad beschriebenen aktuellen Arbeiten enthält.

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

Die Repositories „olena“ bzw. „ocrd_olena“ stellen einen weiterentwickelten Fork des Original-Olena-Repositorys sowie OCR-D-Wrapper für die Olena-Nutzung bereit. Beide Repositories sind in einem stabilen Zustand und werden nach Aussagen der Entwickler nicht mehr weiterentwickelt.

Das Schreiben von Tests wird bei Pull Requests angeregt, aber nicht als zwingende Voraussetzung eingefordert, um potenzielle Beitragende nicht abzuschrecken. Dies entspricht zwar nicht der Vorgehensweise, die als Ziel z.B. beim Kitodo.Production-Repository verfolgt wird, ist aber eine nachvollziehbare Strategie.

Einige Pull Requests sind bereits mehrere Jahre alt. Dies ist zwar für Open Source-Projekte nicht ungewöhnlich. Es sollte aber dennoch in Betracht gezogen werden, Pull Requests, an denen von einem Reviewer formulierte Änderungswünsche nach einer zu definierenden Zeit nicht umgesetzt wurden, temporär zu schließen, um eine Übersicht über aktuelle Tätigkeiten in einem Projekt zu wahren. Diese Pull Requests können zu einem späteren Zeitpunkt gegebenenfalls wieder geöffnet werden, wenn wieder aktiv an ihnen gearbeitet wird.

Ergebnis: Die untersuchten Repositories enthalten sehr unterschiedliche Mengen an offenen bzw. geschlossenen Pull Requests. Während der Code des „core“-Repositories auch aktuell häufig weiterentwickelt wird und das dazugehörige Repository dementsprechend viele offene Pull Requests sowie Issues enthält, ist das „olena“-Repository größtenteils statisch. Dies wird von Aussagen der Entwickler gestützt, die berichten, dass „olena“ nicht mehr weiterentwickelt wird. Ähnliches trifft für die beiden Repositories „ocrd_tesserocr“ und „ocrd_fileformat“ zu, die sich laut Entwicklerangaben in einem stabilen Zustand befinden und für die bis auf zusätzliche Tests derzeit keine Weiterentwicklung geplant ist.
Insgesamt befinden sich die betrachteten Repositories in einem guten Zustand und es konnten keine gravierenden Probleme bzgl. der genannten Aspekte festgestellt werden.

WP 2: Stichprobenhafte Prüfung der Lizenzen

OCR-D steht unter der Apache 2-Open Source-Software-Lizenz. In diesem Arbeitspaket wurde die Kompatibilität der Lizenz des Projektes mit den Lizenzen der verwendeten Abhängigkeiten untersucht.

Um die Kompatibilität von Open Source-Software-Lizenzen zu prüfen, wurde die Wikipedia-Definition unter https://en.wikipedia.org/wiki/License_compatibility zu Rate gezogen und der Open Source-Lizenz-Kompatibilitäts-Checker unter https://joinup.ec.europa.eu/collection/eupl/solution/joinup-licensing-assistant/jla-compatibility-checker verwendet.

Die folgende Liste enthält die Open Source-Lizenzen der in den einzelnen OCR-D-Repositories verwendeten Abhängigkeiten:

-	„atomicwrites“:	MIT	
-	„beanie“:	Apache-2.0	
-	„click“:	BSD-3-Clause	
-	„depcrecated“:	MIT	
-	„docker-py“:	Apache 2.0	
-	„fastapi“:	MIT	
-	„filetype.py“:	MIT	
-	„flask“:	BSD-3-Clause	
-	„frozendict“:	**LGPL-3.0**	X
-	„gdown“:	MIT	
-	„httpx“:	BSD-3-Clause	

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

- „importlib_metadata“: Apache-2.0
- „importlib_resources“: Apache-2.0
- „jsonschema“: MIT
- „lxml“: Copyright license
- „memory_profiler“: BSD
- “numpy”: Copyright licence
- “bagit-profiles-validator”: Public Domain
- “opencv-python”: MIT
- “paramiko”: **LGPL-2.1** X
- “pika”: BSD-3-Clause
- “pillow”: HPND
- “pydantic”: MIT
- “python-magic”: MIT
- “python-multipart”: Apache-2.0
- “pyyml”: MIT
- “requests”: Apache-2.0
- “requests-unixsocket”: Apache-2.0
- “shapely”: BSD-3-Clause
- “uvicorn”: BSD-3-Clause
- “actions/checkout”: MIT
- “docker/login-action”: Apache-2.0
- “docker/setup-buildx-a…”: Apache-2.0
- “codecov-python”: Apache-2.0
- “coveragepy”: Apache-2.0
- “docstr_coverage”: MIT
- “pylint”: **GPL-2.0**
- “pytest”: MIT
- “pytest-benchmark”: BSD-2-Clause
- “recommonmark”: MIT
- “sphinx”: BSD
- “sphinx-click”: Copyright
- “twine”: Apache-2.0
- “types-lxml”: Apache-2.0
- “wheel”: MIT
- “actions/setup-python”: MIT
- “homebrew/setup/” BSD-2-Clause

Die Lizenzen der einzelnen Abhängigkeiten wurden über die jeweiligen GitHub-Repositories ermittelt.

Die Kompatibilität wurde bestimmt, indem die Lizenz der jeweiligen Abhängigkeit als „input licence“ und Apache 2 als „outbound licence“ des OCR-D-Projekts in dem oben erwähnten Licence Checker eingegeben wurden.

Ergebnis: Der Großteil der verwendeten Software-Pakete ist lizenztechnisch mit OCR-D kompatibel. Die Abhängigkeiten, deren Open Source-Lizenzen nicht mit der Apache 2.0-Lizenz des OCR-D-core-Repositories kompatibel sind bzw. weiterer Klärung bedürfen, sind im folgenden aufgelistet:

1. „**frozendict**“ (Lizenz: **LGPL-3.0**)
2. „**paramiko**“ (Lizenz: **LGPL-2.1**)

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

Gemäß des oben erwähnten Lizenz-Checker-Tools dürfen Software-Komponenten, die unter einer LGPL-Lizenz veröffentlicht werden, nur dann in einer unter Apache 2 lizensierten Software verwendet werden, wenn diese ebenfalls unter die gleiche LPGL-Lizenz gestellt wird. Das Ergebnis der Prüfung besagt:

> „Compatibility between the GNU Lesser General Public License (LGPL) 2.1 (inbound licence) and the Apache License, Version 2.0 (outbound licence):
> The inbound licence is a copyleft licence: In case you merge data or source code covered by inbound licence with your work or with any other work, and in case this combination is not only used internally but distributed to third parties, this distribution cannot be done under outbound licence but must be done under inbound licence."
> (Quelle: https://joinup.ec.europa.eu/licence/compatibility-check/LGPL-2.1/Apache-2.0)

Weitere Software-Pakete, wie z.B. „pylint", die unter einer potenziell problematischen GPL-Lizenz stehen, werden nur für die Testausführung verwendet und sind nicht Teil der OCR-D-Distributionen.

Ob die hier festgestellten Inkompatibilitäten in der Praxis ein Risiko darstellen oder nicht kann mangels Expertenwissens auf dem Gebiet Lizenzen nicht abschließend vom Reviewer beurteilt werden und sollte von entsprechenden Fachleuten abschließend geklärt werden.

Falls die genannten Inkompatibilitäten als tatsächliches Problem eingeordnet werden, sollten nach Möglichkeit für die erwähnten Software-Abhängigkeiten entweder Ersatzkomponenten mit kompatiblen Lizenzen gefunden oder das OCR-D-Core-Repository ebenfalls unter eine LGPL-Lizenz gestellt werden.

WP 3: Stichprobenhafte Prüfung der "lines-of-code"

In diesem Arbeitspaket sollte ermittelt werden, ob die Angaben über die Anzahl an Code-Zeilen („lines of code") in den betrachteten Modulen mit den ermittelten Werten übereinstimmen. Zur Ermittlung der tatsächlichen „lines of code" wurde das Plugin „Statistics" der Python IDE „PyCharm" verwendet.

OCR-D/core:

Modul	Angabe in Pad	Tatsächlich ermittelt
ocrd	3863 (py), 201 (sh)	3467 (py), 210 (sh)
ocrd_modelfactory	64 (py)	64 (py)
ocrd_models	1247 (py), 24 (xml)	1792 (py), 24 (xml)
ocrd_network	3513 (py)	3410 (py)
ocrd_page_user_methods	252 (py)	252 (py)
ocrd_utils	1108 (py)	1046 (py), 65 (conf)
ocrd_validators	794 (py)	791 (py), 533 (yml)

OCR-D/ocrd_olena:

Modul	Angabe in Pad	Tatsächlich ermittelt
Top dir	328 (sh)	119 (sh)

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

OCR-D/ocrd_tesserocr:

Modul	Angabe in Pad	Tatsächlich ermittelt
ocrd_tesserocr	2019 (py)	2016 (py), 575 (json)

OCR-D/ocrd_fileformat:

Modul	Angabe in Pad	Tatsächlich ermittelt
Top dir	85 (sh)	408 (sh)

Ergebnis: Die im „Pad“ angegebenen „Lines of Code“ sind größtenteils korrekt. Es konnten nur kleine Abweichungen festgestellt werden, die sich vermutlich aus Änderungen am Code ergeben, die seit der Dokumentation im Pad vorgenommen wurden und sich im für aktive Open Source-Projekte erwartbaren Rahmen bewegen. Bei der Bestimmung der Lines of Code im „core“-Repository wurden die aus XML-Schema-Definitionen automatisch generierten Dateien „ocrd_page_generateds.py“ und „ocrd_page.py“ nicht berücksichtigt.

WP 4: Stichprobenhafte Prüfung der Unit-Test-Abdeckung

In diesem Arbeitspaket sollte die Testabdeckung der betrachteten Repositories untersucht werden. Hierzu wurden im bereitgestellten Pad gemachte Angaben mit selbst ermittelten Werten verglichen.

OCR-D/core:

Im Pad ist die Anzahl der Unit Tests mit 451 angegeben. Die Durchführung der Tests mit „pytest“ ergibt das folgende Ergebnis:

- 20 failed
- 440 passed
- 5 skipped
- 48 warnings
- 34 errors

Dies ergibt insgesamt 547 Tests. Die Testabdeckung ist damit höher als in den bereitgestellten Daten angegeben wurde.

OCR-D/ocrd_olena:

Für das Repository “ocrd_olena” wurden keine Angaben zur Testabdeckung gemacht. Die im Pad erwähnte Datei „test.sh“ ist verständlich geschrieben, nutzt sprechende Variablennamen wie „workspace_dir“ und „binarized_pages“ sowie umfangreiche Logausgaben. Trotz sparsamer Kommentierung des Codes wird daher ersichtlich, welche Funktionalität getestet wird.

OCR-D/ocrd_tesserocr:

Über die Test-Abdeckung des Repositorys „ocrd_tesserocr“ wurde im Pad keine Angabe gemacht. Über das Makefile des Projeks wurde eine Abdeckung von 63% ermittelt.

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

OCR-D/ocrd_fileformat:

Die im Pad angegebene Anzahl von zwei Unit-Tests stimmt mit dem aktuellen Code überein. Die Tests enthalten keine Kommentare, aber die Namen der Testmethoden sind selbsterklärend, daher kann auf weiterführende Erläuterungen verzichtet werden. In einem Test („test_convert.py“) ist ein zum Zeitpunkt des Reviews noch nicht umgesetztes „TODO“ enthalten. Da der dazugehörige Test („test_page_to_alto“) mit der Überprüfung der Transformationsergebnisse eine zentrale Funktionalität des Moduls bzw. Prozessors behandelt, sollte dieses Vorhaben möglichst vor Abschluss des Projektes umgesetzt werden.

Ergebnis: Die bereitgestellten Angaben zur Testabdeckung in den betrachteten Repositories werden im aktuellen Stand des Codes übertroffen. Dies liegt an den Weiterentwicklungen, die seit der Erstellung der zur Verfügung gestellten Dokumentation durchgeführt wurden. Die Repositories „ocrd_olena“, „ocrd_tesserocr“ und „ocrd_fileformat“ enthalten keine Angaben über die Testabdeckung. Die enthaltenen Tests sind gut und verständlich geschrieben. Hier wäre einzig eine ausführliche Dokumentation der Tests durch Kommentare wünschenswert.

Als Nebenbeobachtung soll erwähnt werden, dass die Testausführung im Modul „core“ unter bestimmten Bedingungen fehlschlägt, da Python beim Versuch, die Datei „test_mets_server.py“ auszuführen abstürzt. Wird dieser Test entfernt bzw. übersprungen, laufen die Tests durch.
Nach Rücksprache mit den Entwicklern scheint die Ursache hierfür zu sein, dass der genannte Test nicht mit dem Betriebssystem „MacOS“ kompatibel ist, welches auf dem Rechner läuft, der für das Review verwendet wurde. Die Testausführung sollte jedoch nicht von lokalen Betriebssystemen oder Entwicklungsumgebungen abhängen und alle notwendigen Informationen und Abhängigkeiten mitbringen, um kontextunabhängig erfolgreich durchgeführt werden zu können. Wenn ein Test unter bestimmten Betriebssystemen nicht funktional ist, sollte er unter diesen Umständen automatisch übersprungen werden.

WP 5: Stichprobenhafte Analyse der wesentlichen 3rd Party SW-Pakete

In diesem Arbeitspaket sollte stichprobenhaft geprüft werden, in welchem Zustand sich wesentliche 3rd Party-Software-Pakete befinden, die vom OCR-D-Projekt bzw. den zu untersuchenden Repositories verwendet werden.

OCR-D/core:

Als wesentliche 3rd Party-Software-Komponenten wurden die folgenden im Pad dokumentierten Abhängigkeiten gewählt und betrachtet:

1. „requests“:
    i. letzter Release: 2.31.0, vom 22.05.2023
    ii. damit als aktuelle Software zu betrachten (neuester Release nicht älter als 1 Jahr)
    iii. Major Version „2“ impliziert Ausgereiftheit
    iv. Core nutzt bis maximal Version 2.29, vom 26.04.2023
    v. Evtl. Upgrade auf aktuelle Version möglich, da nur Bugfix release? Laut Release Notes enthält diese allerdings potenziell „breaking changes“
2. „fastapi“:
    i. letzter Release: 0.111.0, vom 03.05.2024
    ii. Als sehr aktuelle Software zu betrachten (neuester Release weniger als eine Woche alt)

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

iii. Allerdings noch in einem potenziell nicht stabilen Zustand (Major Version „0“ impliziert frühen Entwicklungsstatus)
iv. Erste Version allerdings bereits 6 Jahre alt, daher vielleicht ausgereifter als Versionsnummer vermuten lässt
v. „Core“ nutzt Version 0.78.0 (oder höher) vom 14.05.2022;
vi. Releases sehr ausführlich dokumentiert

3. „pika“:
   i. Aktuellster Release: 1.3.2, vom 05.05.2023
   ii. Damit als aktuelle Software zu betrachten (neuester Release ein Jahr alt)
   iii. Major Version „1“ impliziert Ausgereiftheit
   iv. „core“ nutzt Version 1.2.0 (vom 05.02.2021) oder höher; da hier auch aktuelle Versionen genutzt werden können, wird kein Handlungsbedarf gesehen
4. „beanie“:
   i. Aktuellster Release: 1.26.0, von Mai 2024
   ii. Damit als sehr aktuelle Software zu betrachten, an der aktiv entwickelt wird
   iii. Major Version-Nummer „1“ impliziert Ausgereiftheit
   iv. „core“ nutzt Version 1.7 (vom 29.11.2021);
5. „numpy“:
   i. Aktuellster Release: 1.26.4, vom 06.02.2024
   ii. Weit verbreitete, ausgereifte Bibliothek für numerische Berechnungen mit Pyhton
   iii. Keine Versionsvorgabe (um Kompatibilität mit verschiedenen „tensorflow“-Versionen zu gewährleisten)

OCR-D/ocrd_olena:

„ocrd_olena“ befindet sich gemäß bereitgestellten Informationen in einem stabilen Zustand und es sind keine weiteren Arbeiten daran geplant. Die einzige, wesentliche Abhängigkeit, die ermittelt werden konnte, ist die in C++ geschriebene Bildbearbeitungsbibliothek Olena. Diese befindet sich laut Aussagen der Entwickler in einem stabilen Zustand, wurde jedoch seit mehreren Jahren nicht mehr von den Originalautoren weiterentwickelt. Der letzte Commit im dazugehörigen Repository https://github.com/glazzara/olena ist 10 Jahre alt. Die in diesem Repository erwähnte Dokumentations-Seite http://olena.lrde.epita.fr/ ist zum Zeitpunkt dieses Reviews nicht erreichbar bzw. liefert eine 404-Fehlermeldung zurück.

Es erscheint fraglich, ob gefundene Bugs oder Sicherheitslücken in diesem Projekt bzw. Repository noch behoben werden. Daher wird diese Aufgabe vermutlich auf Projekte fallen, die das Olena-Repository verwenden. Dies wird im OCR-D-Kontext scheinbar bereits gemacht (https://github.com/glazzara/olena/compare/master...OCR-D:olena:master). Die Entwickler der OCR-D-Repositories haben bestätigt, dass Anpassungsarbeiten und Weiterentwicklungen vorgenommen wurden, um Olena im OCR-D-Kontext nutzbar zu machen.

OCR-D/ocrd_tesserocr:

1. „tesseract“:
   i. Aktuellster Release: 5.3.4, vom 18.01.2024
   ii. Major Versions-Nummer „5“ impliziert einen sehr ausgereiften Status

OCR-D/ocrd_fileformat:

- Es konnten keine wesentlichen Abhängigkeiten außerhalb des OCR-D-Kontexts ermittelt werden

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

Ergebnis: Nach den ermittelten Informationen befindet sich der größte Teil der verwendeten 3rd Software-Komponenten in einem ausgereiften Zustand, wird jedoch weiterhin aktiv weiterentwickelt und gewartet. Eine Ausnahme stellt die Bildbearbeitungsbibliothek „Olena“ dar, deren Original-Repository seit längerer Zeit nicht mehr aktiv ist. Daher fallen die Wartung und Weiterentwicklung des Codes für das OCR-D-Projekt in diesem Fall auf dessen Entwickler zurück. Derzeit sind nach Entwickleraussagen jedoch keine weiteren Arbeiten an Olena geplant.

WP 6: Stichprobenhafte Prüfung der Coding-Guidelines

OCR-D/core:

Die im Pad erwähnte flake8-Konfigurationsdatei ist nicht teil des Repositories und stand nicht für die Überprüfung der Coding Guidelines zur Verfügung. Laut Entwickler-Aussagen ist kann die Einhaltung der für das OCR-D-Projekt angepassten PEP8-Regeln jedoch mit dem folgenden parametrisierten Kommandozeilenaufruf des flake8-Tools überprüft werden:

```
flake8 --select E,F,B,Q,ICN,TID --ignore E501,F841,Q000,E30,E11,E12,E2,F54
--extend-exclude 'ocrd_page_*' src
```

Diese Parameter werden daher als Grundlage der Coding Guidelines des OCR-D-Projektes verwendet und überprüft.

Mithilfe dieses Aufrufs wurden insgesamt 210 Fehler bzgl. angepassten PEP8-Regeln festgestellt. Diese teilen sich wie folgt auf die folgenden Bereiche auf:

- E402 (module level import not at top of file): 14
- E502 (the backslash is redundant between brackets): 6
- E701 (multiple statements on one line (colon)): 10
- E713 (test for membership should be ‘not in’): 16
- F401 (‘ocrd.processor.base.run_processor’ imported but not used): 160
- F403 (‘from ocrd_validators import *’ used; unable to detect undefined names): 1
- F507 ('...' % ... has 4 placeholder(s) but 3 substitution(s)): 1
- F811 (redefinition of unused 'ET' from line 6): 1
- F821 (undefined name 'scale_coordinates'): 1

Die Entwickler haben erläutert, dass es sich bei den meisten der detektierten Fehler vom Typ „F401“ um „false positives“ handelt, die in den Paketdefinitionsdateien __init__.py auftauchen. Diese Aussage konnte durch das Review bestätigt werden. Werden diese vermeintlichen Import-Fehler aus der Statistik entfernt, bleiben nur noch 31 tatsächliche ungenutzte Importe im Quellcode. Die Gesamtzahl der gefunden Probleme reduziert sich damit von 210 auf 81.

Es ist nicht klar definiert, wie die Erfüllung von Coding Guidelines quantifiziert wird. Die Anzahl der durch das automatische Tool „flake8“ gefundenen Mängel bzgl. der in PEP8 definierten Coding Guidelines muss in einen Zusammenhang mit dem gesamten Code-Umfang gebracht werden. Die Python-Dateien im „core“-Modul von OCR-D enthalten insgesamt 10852 Zeilen Code (siehe Ergebnis WP 3)

Wird der parametrisierte Aufruf von flake8 verwendet, ergibt sich, dass
81 von 10852 Zeilen mindestens eine Coding Guideline von OCR-D verletzen, was einer Quote von 0,75% problembehafteten Code-Zeilen entspricht. Dies zeigt, dass der Code des „core“-Repositories bzgl. der

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

selbstauferlegten Coding Guidelines sehr bereinigt und in einem guten Zustand ist. Es ist jedoch zu beachten, dass der parametrisierte Aufruf einige PEP8-Regeln explizit ignoriert.

Des Weiteren ist die erwähnte Quote nur eingeschränkt aussagekräftig, da es sein kann, dass sich in einer Zeile mehrere Fehler befinden, was aus der obigen Analyse nicht hervorgeht.

Die meisten der detektierten Probleme sind leicht zu beheben, z.B. indem mehrere durch Semikolon getrennte Anweisungen in einer Zeile in mehrere Zeilen aufgeteilt werden.
Da die Behebung der meisten gefundenen Mängel trivial ist, stellt die Anzahl der gefundenen Probleme nach Einschätzung des Reviewers kein großes Problem dar. Das verantwortliche Team sollte jedoch evaluieren, ob die aktuellen Coding Guidelines in Form des erwähnten, parametrisierten Aufrufs des flake8-Tools weiterverwendet werden sollen oder eventuell mehr PEP8-Regeln berücksichtigt werden können.

OCR-D/ocrd_olena:

Es wurden keine Angaben zu konkreten Coding Guidelines gefunden. Ein Test mit dem Tool „shellcheck“ (www.shellcheck.net) auf den entsprechenden Dateien ergab keine signifikanten Code Style-Probleme.

OCR-D/ocrd_tesserocr:

Mithilfe des parametrisierten flake8-Aufrufs (siehe oben, Abschnitt „core“) konnten keine Probleme festgestellt werden. Es wurden lediglich acht vermeintlich ungenutzte Import-Statements gefunden, bei denen es sich jedoch wie im Falle des „core“-Repositorys um „false positives“ handelt.

OCR-D/ocrd_fileformat:

Es wurden keine Angaben zu konkreten Coding Guidelines gefunden. Ein Test mit dem Tool „shellcheck“ (www.shellcheck.net) auf den entsprechenden Dateien ergab keine signifikanten Code Style-Probleme.

Ergebnis: Die Coding Guidelines für Python-Dateien richten sich nach den weit verbreiteten PEP8-Regeln und werden in den betrachteten Repositories größtenteils eingehalten. Da jedoch eine nicht zu vernachlässigende Menge von PEP8-Regeln bei den Coding Guidelines nicht betrachtet wird, kann diesem Ergebnis nur eine eingeschränkte Aussagekraft zugeordnet werden. Des Weiteren wäre es wünschenswert, die Coding Guidelines öffentlich in einem entsprechenden Dokument zu erfassen, damit Entwickler wissen, welche Regeln beim Beitragen von Code eingehalten werden sollen.

WP 7: Analyse zu Security-Aspekten

In diesem Arbeitspaket soll beurteilt werden, in welchem Ausmaß „Security“-Aspekte beim Betrieb des OCR-D-Frameworks relevant sind und beachtet werden müssen.

Prinzipiell wird der Annahme des Auftraggebers zugestimmt, dass dem Thema „Security“ im Kontext von OCR-D im Vergleich zu anderen Web-Applikationen eine untergeordnete Rolle zukommen kann. Zum einen erlaubt das Betreiben der Komponenten in Docker-Containern eine saubere Trennung von Docker- und Host-Systemen, die eine hohe Abschottung der Systeme und damit eine hohe Sicherheit gewährleistet. Zum anderen gewährleistet der Einsatz in lokalen Umgebungen, die durch Firewalls geschützt sind, eine zusätzliche Absicherung.

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

Trotz der generell hohen Sicherheit von in Docker-Containern betriebenen Systemen sollte darauf geachtet werden, die verwendeten Docker-Images aktuell zu halten. In diesem Zusammenhang sollte eine Roadmap existieren, zu welchem Zeitpunkt beispielsweise das Docker-Image des im „core“-Repository verwendeten „Ubuntu“-Betriebssystems von Ubuntu 20 auf die derzeit aktuelle Version 22 aktualisiert wird.

WP 8: Besondere Analyse „Web-Interface“

Das Web-Interface ermöglicht das Anstoßen von OCR-D-Funktionalitäten über standardisierte REST-Endpunkte. Die zur Verfügung stehenden Endpunkte ermöglichen die Durchführung spezifischer OCR-D-Funktionalitäten entweder auf Prozessor- oder auf Workflow-Ebene mit eigenen Daten in Form von sogenannten „Workspaces“.

Der Stand und die Verwendung des Web-Interfaces, wird auf https://ocr-d.de/en/spec/web_api gut dokumentiert. Hier wäre eine deutsche Übersetzung der Dokumentation wünschenswert.

Die Entwicklung des Moduls ist nach Angaben der Entwickler weitestgehend abgeschlossen und das Web-Interface kann damit als bereit für den Produktiveinsatz betrachtet werden. Es stehen keine großen Weiterentwicklungen an.

Im Modul „core/tests/network“ sind ausführliche Tests für das Web-Interface vorhanden, die nach einer Stichprobenhaften Analyse einen großen Teil der verfügbaren REST-Endpunkte bzw. die dahinterstehende Funktionalität testen. Damit können Sie als Nachweis der generellen Funktionsfähigkeit des Web-Inferface angesehen werden.

Ergebnis: Das Web-Interface von OCR-D kann als ausgereift und einsatzbereit betrachtet werden. Die Dokumentation enthält alle wichtigen Informationen zur Verwendung des Interfaces, sollte im Idealfall aber noch übersetzt werden.

WP 9: Dokumentation

In diesem Arbeitspaket wurde die Dokumentation des OCR-D-Projektes betrachtet.

Release Management:

- Der Absatz zur Dokumentation für CI/CD beschreibt die automatisch durchgeführten Tests und Build-Prozesse.
- Eine explizite und öffentlich einsehbare Dokumentation der angesetzten Coding Guidelines fehlt und sollte zur Verfügung gestellt werden.

Produktions-/Betreiber-Dokumentation:

- Die Betreiber-Dokumentation ist ausreichend dokumentiert. Die Möglichkeit, das System als Docker-Container aufsetzen zu können, wird ausdrücklich begrüßt und ist sehr ausführlich.

Management der Dokumentationsseiten:

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

- Die Dokumentationsseiten werden über das GitHub-Repository https://github.com/OCR-D/ocrd-website/wiki organisiert. Änderungen und Korrekturen können über Pull Request von der Community vorgenommen werden.

Die meisten Seiten der Dokumentation stehen sowohl auf Deutsch als auch auf Englisch zur Verfügung. Es gibt jedoch Ausnahmen (z.B. https://ocr-d.de/en/spec/web_api), von denen derzeit nur eine englische Variante existiert. Es sollte angestrebt werden, alle Seiten in beiden Sprachen anzubieten.

Eine Unterscheidung nach den als „minimal“ beschriebenen Paketen wurde bei der Beurteilung der Dokumentation nicht vorgenommen, da eine Betriebsdokumentation für Module wie „ocrd_olena“ nicht sinnvoll erscheint. Diese Module werden nicht allein in Betrieb genommen, sondern müssen im Kontext eine kompletten OCR-D-Installation verstanden und damit auch dokumentiert werden. Dies ist nach unserer Meinung gegeben.

Ergebnis: Insgesamt wird die Dokumentation der betrachteten OCR-D-Repositories als sehr ausführlich und umfangreich wahrgenommen. Die verschiedenen Möglichkeiten der Installation (Docker vs Native) sind sehr verständlich dokumentiert.

WP 10: Stichprobenhafte Analyse der Code-Qualität

Bei der stichprobenhaften Analyse der Code-Qualität wurde sowohl die Einhaltung etablierter Best Practices in der Programmierung wie maximale Zeilen- und Methodenlängen als auch die Berücksichtigung bestimmter Programmier-Paradigmen untersucht, die speziell bei der Python-Programmierung als „Best Practices“ betrachtet und als „Pythonic“ bezeichnet werden. Hierzu gehört die Verwendung von „for“-Schleifen ohne Zähler-Variablen, die Erstellung von Listen mit sogenannten „List Comprehensions“ und die Verfolgung der Philosophie „asking for forgiveness is easier than asking for permission“, die die Verwendung von „try“-“catch“-Blöcken als Alternative zu „if“-„else“-Konstrukten anregt. Hierbei wurden jeweils mehrere Klassen aus den Repositories „core“ und „tesserocr“ untersucht. Die Repositories „olena“ und „ocrd_fileformat“ wurden hierbei übersprungen, da sie nur ein paar Shell-Skripte enthielten und keine umfangreichen Code-Pakete bzw. -Module.

Die folgenden Punkte wurden konkret überprüft.

1. Allgemeine „Best Practices“:
    a. sind alle „public“ Methoden mit einem Kommentar versehen? Als „public“ Methoden gelten in Python solche, deren Funktionsname nicht mit einem „_“ oder „__“ beginnt
    b. sind alle „public“ Methoden höchstens 50 Zeilen lang?
    c. Strukturierung, d.h. „separation of concerns”; Klassengröße;

2. Python-spezifische „Best Practices“:
    a. Verwendung von „List Comprehensions“ und „Dict Comprehensions“
    b. „Pythonic Tuple“ zum Entpacken mehrerer Elemente aus einer Liste in einem Schritt
    c. Datei-Handling mit „open()“

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

d. Verwendung von „try“-„catch“-Blöcken wo sinnvoll
e. Verwendung der „enumerate“-Funktion zum Iterieren über iterierbaren Objekten, bei denen auch eine Index-Variable benötigt wird

Die Überprüfung dieser Kriterien überschneidet sich teilweise mit der Überprüfung der Einhaltung von Coding Guidelines aus WP 6, geht aber anhand von einzelnen Klassen mehr ins Detail.

OCR-D/core:

d. `src/ocrd/cli/workspace.py`:
    i. umgesetzte Python-Idiome:
        1. „for“-Schleifen ohne Zähler-Variablen
        2. „asking for forgiveness is easier than asking for permission“: use „try”…”catch” instead of “if”…”else”
        3. Einhaltung der “one statement per line“-Regel (keine „;“ zum Gruppieren mehrerer Statements!)
        4. Alle „public“ (d.h. ohne „_“ oder „__“) Methoden enthalten einen Kommentar
        5. 25/26 Methoden haben höchstens 50 Zeilen
    ii. Einschränkungen/Probleme:
        1. 1/26 Methoden hat mehr als 50 Zeilen Länge: „workspace_cli_bulk_add“: 142 Zeilen
e. `src/ocrd/processor/base.py`:
    i. umgesetzte Python-Idiome:
        1. „for“-Schleifen ohne Zähler-Variablen
        2. „asking for forgiveness is easier than asking for permission“: use „try”…”catch” instead of “if”…”else”
        3. Einhaltung der “one statement per line“-Regel (keine „;“ zum Gruppieren mehrerer Statements!)
        4. 9/11 „public“-Methoden und -Properties sind kommentiert;
        5. 10/11 „public“-Methoden haben weniger als 50 Zeilen
    ii.
        1. 2/11 Methoden ohne Kommentar („Show_help“ und „show_version“)
        2. 1/11 Methoden hat mehr Zeilen („zip_input_files“)
f. `src/ocrd_network/utils.py`:
    i. umgesetzte Python-Idiome:
        1. „for“-Schleifen ohne Zähler-Variablen
        2.

OCR-D/ocrd_olena:

- Nur Shell-Skripte, keine nähere Analyse

OCR-D/ocrd_tesserocr:

- Keine nähere Analyse

OCR-D/ocrd_fileformat:

- Nur Shell-Skripte, keine nähere Analyse

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23

Ergebnis: In den betrachteten Repositories werden sowohl allgemeine als auch Python-spezifische „best practices“ in den untersuchten Code-Stichproben umgesetzt. Dazu gehören Methodenlängen und kommentierte, öffentliche Methoden einerseits sowie die Verwendung von Ausnahmenbehandlung statt Fallunterscheidung und „for“-Schleifen ohne Index-Variablen andererseits. Verletzungen dieser Regeln konnten in den untersuchten Code-Stichproben nur vereinzelt festgestellt werden.

WP 11: Subjektive Einschätzung des Auftragnehmers

Die in diesem Review betrachteten Repositories „core“, „ocrd_olena“, „ocrd_tesserocr“ und „ocrd_fileformat“ erscheinen in einem sehr ausgereiften und gut gepflegten Zustand. Die komplexe Projekt-Struktur wird durch die zur Verfügung stehenden Dokumentationen und Diagramme sowohl auf der OCR-D-Webseite (https://ocr-d.de) als auch im Wiki des GitHub-Repositories (https://github.com/OCR-D/ocrd-website/wiki) verständlich erklärt. Die Community bzw. die Entwickler reagieren schnell und hilfsbereit auf Fragen und helfen bei Problemen weiter. Dies ist nach Ansicht des Reviewers für Open Source-Projekte essenziell.

Die Testabdeckung der Repositories ist trotz mangelnder Verpflichtung zum Schreiben von Tests für neue Funktionen und Module hoch, wodurch erfolgreich durchgeführten Builds während der CI eine höhere Bedeutung zukommt als bei Software-Projekten mit niedriger Testabdeckung.

Aufgrund der Komplexität und Größe des gesamten OCR-D-Projektes kann die Aussage der betrachteten Repositories jedoch nicht ohne separate Analyse auf alle anderen Projektteile erweitert werden. Da die verschiedenen Teilprojekte von OCR-D laut Aussage der Entwickler getrennt und von unterschiedlichen Einrichtungen entwickelt wurden, kann daher von dem Zustand eines Repositorys nicht auf den Zustand anderer Repositories gefolgert werden.

Die Qualität der betrachteten Repositories wird abschließend bzgl. der untersuchten Faktoren als hoch und ausgereift eingestuft und birgt nach unserer Einschätzung kein erhöhtes Risiko für ein zukünftiges neues Release Management.

Arved Solth, 17.06.2024

effective WEBWORK GmbH
Neuer Wall 18
DE-20354 Hamburg

Tel.: +49(0) 40/60 94 08 57-0
Fax: +49(0) 40/60 94 08 57-1
E-Mail: info@effective-webwork.de

Geschäftsführung: Anne Christensen,
Prof. Dr. Matthias Finck, Josephine Kraus
Handelsregister: HRB 101108 (Hamburg)
Umsatzsteuer-ID-Nr.: DE254169943

Ethikbank
BIC: GENODEF1ETK
IBAN: DE27 8309 4495 0003 3523 23