Open Access Indicator for 2019

Part 2

Technical Description of Data Foundation, Processes and Output Publications from publication year 2017

0   Preface

Denmark's National Strategy for Open Access 2018-2025[1] has the target: *"From 2025 onwards, there should be unimpeded digital access for all to all peer-reviewed scientific articles from Danish research institutions – with a maximum 12- month embargo."*

Furthermore, the strategy specifies: *"Once a year, the Ministry of Higher Education and Science will monitor the implementation of Open Access via the Open Access Indicator (OAI) to support that all parties are doing their utmost to develop and expand free accessibility to Danish research results."*

The Open Access Indicator is calculated for each publication year with the scope/target field set to: "*Scientific articles and conference contributions in journals and proceedings with an ISSN number.*

Many journals maintain embargo periods of 6, 12, or more months, during which researchers are blocked from establishing open Access to their articles. Similar to the EU Horizon 2020[2] research programme, the national strategy allows up to 12 months delay in establishing Open Access to a given article.

The Open Access Indicator (OAI) is calculated once annually with a delay that allows for a potential 12-month embargo period for articles published at the end of a publication year. Thus OAI 2019 is calculated early March 2019 for publications from the publication year 2017.

The description of the Open Access Indicator is organized in two parts:

- Part 1: Overview of data foundation, processes and output
- Part 2: Technical description of data foundation, processes and output

   Note: In Part 2, the technical description, the notion of the indicator's "target field" is expressed using the term "set of scoped records".

Queries regarding the indicator may be directed to

Hanne-Louise Kirkegaard
Danish Agency for Science and Higher Education
Ministry of Higher Education and Science
Bredgade 40
DK-1260 København K
Email: hki@ufm.dk

---

[1] https://ufm.dk/en/research-and-innovation/cooperation-between-research-and-innovation/open-access/Publications/denmarks-national-strategy-for-open-access/denmarks-national-strategy-for-open-access
[2] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

1   Introduction and Main Processes

The activities of the OA Indicator can be broken down into these five main processes.
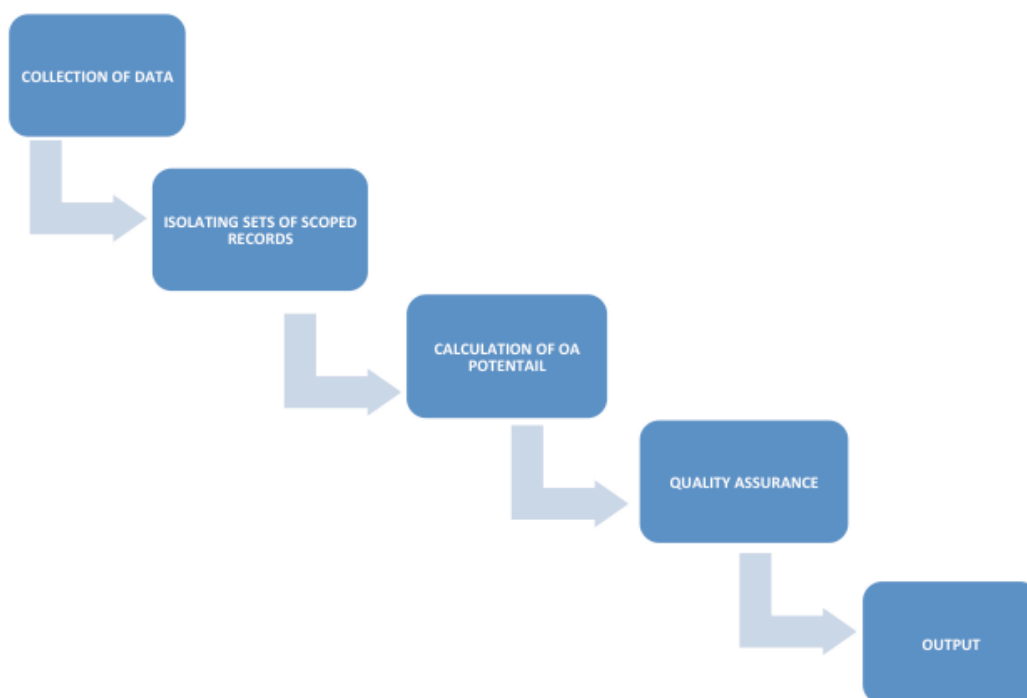
The five main processes are described in further detail in the sections below.

This description of the Open Access Indicator is aimed for a technically inclined audience and aims to describe in depth how the Indicator works – overall as well as in detail.

The description assumes that the reader has familiarity with basic XML[3] and basic parts of the XPath[4] notation for referin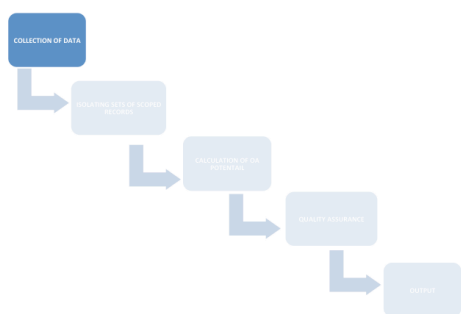g to XML elements of an XML document conforming to a certain XML Schema. It also assumes that the reader is familiar with visualisation of processes af workflow diagrams[5].

---

[3] https://www.w3.org/TR/xml/

[4] https://www.w3.org/TR/xpath-30/

[5] https://en.wikipedia.org/wiki/Flowchart

2   Process 1: Collection of The Data

The first activity in the OA Indicator is the collection of the complete data foundation used by the indicator. This includes importing six national and international sources. The data foundation is composed of metadata describing the publications of the universities, as well as authority- and auxiliary data.

2.1   The Universities Publication Data

Metadata describing the publications of the universities are used to establish the set of publications in scope of the OA Indicator.

Metadata describing the publications of the universities are collected for the OA Indicator once annually. Collection is done directly from the universities, using an XML-based nationally agreed exchange format and a nationally agreed exchange protocol.

For fulltexts registered in the collected publication metadata, collection (download) are attempted.

2.1.1   Requirements on Universities – Metadata Format and Method of Collection

A university can be included in the OA Indicator if it meets the following minimum requirements:

- Publications published by researchers employed at the university are collected in a university research database containing publication data, person data, project data etc of that particular university only.
- This research database of the university must expose its publication data using OAI-PMH (http://www.openarchives.org/OAI/openarchivesprotocol.html).
- The research database must support OAI-PMH *selective harvesting* using *Sets*, characterised by their *setSpec* (code), to harvest only parts of the database.
- A dedicated OAI-PMH Set exposing all publication data held in the research database must exist.
- For this dedicated set, OAI-PMH metdataPrefix "ddf_mxd" must be supported.
- When an OAI-PMH client harvest this dedicated set using metadataPrefix "ddf_mxd", metadata records must be valid DDF-MXD (http://mx.forskningsdatabasen.dk/mxd/).

2.1.2   This Years Universities and Their Research Databases

The following 8 universities – and associated research databases – are included in the OA Indicator for 2017 publications:

University	Research Database - OAI-PMH server	OAI-PMH setSpec
AAU	http://vbn.aau.dk/ws/oai	publications:all
AU	https://pure.au.dk/ws/oai	publications:all
CBS	http://research.cbs.dk/ws/oai	publications:all
DTU	http://orbit.dtu.dk/ws/oai	publications:all
ITU	https://pure.itu.dk/ws/oai	publications:all
KU	http://curis.ku.dk/ws/oai	publications:all
RUC	http://rucforsk.ruc.dk/ws/oai	publications:all
SDU	http://heinz.sdu.dk:8080/ws/oai	publications:all

2.2   Authority and Auxiliary Data

Authority and Auxiliary Data are collected for the OA Indicator from various sources. For each of these sources, the collection is done once annually. Collection method and data formats vary across sources.

2.2.1   Directory of Open Access Journals (DOAJ)

DOAJ is used by the OA Indicator as an authorative list of Golden Open Access Journals as well as the source of data describing if the journal require APC charges or not.
Parameters of the data collection:

- Protocol: OAI-PMH (server http://www.doaj.org/oai/)
- metadataPrefix: oai_dc
- Dataformat: Dublin Core (http://dublincore.org/documents/dces/)
- Enrichment: Per-journal lookup using REST API endpoint https://doaj.org/api/v1/journals (cf. https://doaj.org/api/v1/docs#!/CRUD_Journals/get_api_v1_journals_journal_id)
- Dataformat: JSON

2.2.2   Sherpa/Romeo (Sh/Ro)

Sh/Ro is used by the OA Indicator to determine the policy for Green Open Access by journals, and thereby the Open Access potential of individual journal articles.
Parameters of the data collection:

- Protocol: HTTP (GET from http://www.sherpa.ac.uk/downloads/)
- Dataformat: Proprietary XML-based format (http://sherpa.ac.uk/news/2012-10-08-RoMEO-API-News.html)

2.2.3   The Danish Bibliometric Research Indicator (BFI)

Data from BFI are used by the OA Indicator for two purposes:

- To identify duplicate publication data across universities (exists for collaborative publications with coauthors employed at different universities and therefore registered in multiple research databases)
- To resolve potential conflicts wrt. Main Research Areas registered in the metadata for the publications

Parameters of the data collection:

- Protocol: HTTPS (GET from https://bfi.fi.dk/AnnualReport)
- Format: Compressed Excel spreadsheet – undocumented template

2.2.4 Authority List: Accepted External Repositories ("The Whitelist")
For fulltexts deposited in external repositories, this authority list is used by the OA Indicator to only allow fulltexts deposited in accepted external repositories to demonstrate Realised Open Access Potential.
- Protocol: Mail (from Authority list maintainers)
- Format: Excel Spreadsheet – undocumented template

2.2.5 Authority List: Journals with extended Embargo ("The Blacklist")
The authority list is used by the OA Indicator to reclassify from Unused to blocked Open Access Potential for journals registered on the list.
- Protocol: Mail (from Authority list maintainers)
- Format: Excel Spreadsheet – undocumented template

2.3 This Years Complete Data Collection
Summary of the data collection for the OA Indicator for 2017 publications:

Source	Protocol	Ver.	Format	Ver.	Collection Date
AAU	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
AU	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
CBS	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
DTU	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
ITU	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
KU	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
RUC	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
SDU	OAI-PMH	2.0	DDF-MXD	1.4.0	4/3 – 2019
DOAJ	OAI-PMH	2.0	DC+JSON	%	4/3 – 2019
Sh/Ro	HTTP	%	Proprietary	%	4/3 – 2019
BFI	HTTPS	%	Proprietary	%	17/10 - 2018
Whitelist	Mail	%	Proprietary	%	25/10 - 2018
Blacklist	Mail	%	Proprietary	%	25/1 - 2019

3 Process 2: Defining the Set of In-Scoped Publications
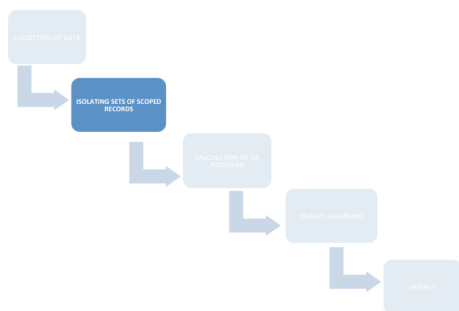
After the collection of all data for the OA Indicator, a number of activities are initiated in order to isolate the publication records which are in scope for the OA Indicator. Not all publications are in scope – only a subset of the publications of the universities.

The scope is defined as:
- *Scientific, peer-reviewed articles and conference contributions published in journals or proceedings with ISSN*

Thus, the subset of publication metadata records representing this scope must be isolated from the total set of publication metadata collected. Ths is done in two ways, in order to facilitate statistics on the national level and on the university level:
- **Scoped records including duplicates – for statistics on the university level**
  For collaborative articles across universities, all registrations from all participating universities are kept
- **Scoped records excluding duplicates – for statistics on the national level**
  For collaborative articles across universities, only one registration is kept.

3.1   The Set of Scoped Records Including Duplicates

Each of the requirements in the definition of the scope maps nicely to a corresponding rule regarding DDF-MXD data elements and their content.

The set of scoped publication metadata records are therefore the set that complies to all the rules. The rules are described below.

First of all, the set of scoped records must represent records with a given submission year. Initial rule is therefore:

0) The *submission year (indberetningsår)* must be marked up in the publication metadata record with the given value.
   Rule applied: Attribute /ddf_doc/@doc_year have the value (year) for the OA indicator calculation

Subsequently, the following four rules are applied on all records:
1) The *type* of the publication must be marked up in the publication metadata record as "Journal Article" "Review article" or "Conference Contribution" (same definition of "article" as used by BFI).
   Rule applied: Attribute /ddf_doc/@doc_type has value "dja", "djr" or "dcp".
2) The *review-status* of the publication must be marked up in the publication metadata record as "Peer-review" (similar demand as for BFI).
   Rule applied: Attribute /ddf_doc/@doc_review has value "pr".
3) The *scientific level* of the publication must be marked up in the publication metadata record as "Scientific" (similar demand as for BFI).
   Rule applied: Attribute /ddf_doc/@doc_level has value "sci"
4) The *publication status* of the publication must be marked up in the publication metadata record as "Published" (similar demand as for BFI).
   Rule applied: Attribute /ddf_doc/publication/*/@pub_status has value "p"

5) The ***publication channel*** of the publication must be marked up in the publication metadata record ***with an ISSN***.
   Rule applied: Element /ddf_doc/publication/*/issn has value.

3.2   The Set of Scoped Records Excluding Duplicates

For collaborative publications between the universities, multiple publication metadata records may represent the same publication. As this is impractical when producing statistics on the national level, a set of scoped records without duplicates are produced.

This set is produced by exposing the set of scoped records with duplicates to a deduplication process. The ambition of this process is to ensure, that for each publication in the scope of the OA Indicator and for which there is *at least one* record in the set of scoped records including duplicates, there is *exactly one* record in the set of scoped records excluding duplicates.

The deduplication proces creates clusters of records. A cluster contains records that represents the same publication. The full set of scoped records excluding duplicates is ultimately established by producing one record per cluster.

The algorithm for producing clusters is:
  1) Records that were part of the BFI calculation for the same submission year and were identified by the BFI process as being duplicates, are added to the same cluster
  2) Records for which significant metadata elements (DOI, title, sub title, ISSN, publication year, etc.) matches sufficiently well, are considered to represent the same publication and are added to the same cluster

This algorithm respects BFI's deduplication algorithm: Rule (1) ensures that any records identified by BFI as duplicates are also identified by the OA Indicator as duplicates.

The scope of BFI and the scope of the OA Indicator differ. This makes it realistic that other non-BFI-scoped records are part of the OA Indicator scope and are indeed duplicates to other records. Rule (2) ensures, that these records are in fact (best effort) being fathomed into clusters as well.

Thus, clusters may include
  a. Only records which were part of BFI,
  b. Both records which were part of BFI and records which were not, or
  c. Only records which were not part of BFI.

A subtle but important remark: For clusters containing BFI records - (a) and (b) above – the BFI records clustered by rule (2) above may stem from different BFI clusters. OA Indicator clusters may contain BFI records which were not joined by the BFI deduplication algorithm.

Conflict Resolution
The results of the OA Indicator are distributed on Main Research Area (MRA). In order to be able to do this distribution, each cluster must have a unique Main Research Area.

BFI's definition of MRA is used by the OA Indicator:
- Science (sci)
- Social Science (soc)
- Humanities (hum)
- Medicine (med)

All DDF-MXD records contain a unique MRA.

For records in the set of scoped records including duplicates, these MRA's are used.

For records in the set of scoped records excluding duplicates, records in the underlying clusters may disagree on MRA. Using BFI terminology, such a situation is called an MRA-conflict. Such MRA-conflicts must be resolved so each cluster has a unique MRA.

The algorithm for resolving MRA-conflicts in a cluster are:
1) If all the records in a cluster have the same MRA, this is used for the cluster (no conflict)
2) Otherwise, if one or more of the records in the cluster were part of a BFI cluster, the BFI MRA for that cluster is used.
3) If none of the records in the cluster were part of the BFI calculation – or if multiple records were part of different BFI clusters diagreeing on their BFI MRA for those BFI-clusters – majority wins: The MRA of the cluster is the MRA represented by most of the records in the cluster.
4) If two or more MRA's are represented by the same number of records in the cluster, the MRA with the highest representation in the entire set of scoped records is chosen for the cluster.

This algorithm ensures, that the OA Indicator solves potential MRA-conflicts respecting to the largest extend possible the corresponding MRA-conflict resolutions done by BFI.

4   Process 3: Calculation of OA Realization and Potential
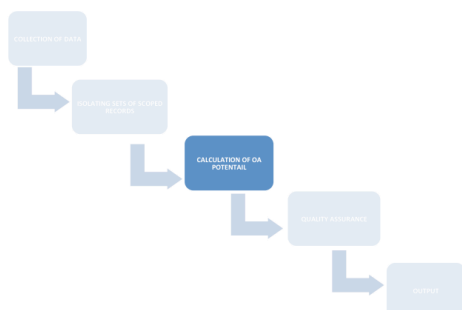
The calculation of OA realisation and potential are done respecting Green and Golden Open Access. The calculation is done nationally, distributed on Main Research Area (MRA) and distributed on universities.

The Open Access potential – and the realisation of that – is initially calculated per university, using a per-publication approach based on the set of scoped records including duplicates. Subsequently, it is also calculated for the national level and MRA level, also using a per-publication approach, but based on the set of scoped records excluding duplicates

For both sets, each record/publication belonging to the set is classified according to how the publication realise its Open Access potential.

There are three values for this classifications, and they are color coded using green, yellow and red (traffic light):
- **Realised** Open Access potential
- **Unused** Open Access potential, and
- **Blocked** Open Access potential

For some in-scoped records, the classification includes attempting a download of a fulltext registered in the record. For technical reasons, the actual download attempts of all potential fulltexts are the first sub process. Please refer to Appendix A for technical details on how this is done.

For records/publications classified as Realised, the types of realisation are also determined. There are four types of Realised:
- Golden Open Access in journals with APC
- Golden Open Access in journals without APC
- Green Open Access from local repository
- Green Open Access from external repository

Each record/publication may have more than one type of realisation.

4.1 Open Access Classification – University Level

For any record in the set of scoped records including duplicates, the Open Access potential is established through a number of validation steps.

As an overview, the classification proces scan be illustrated as follows:
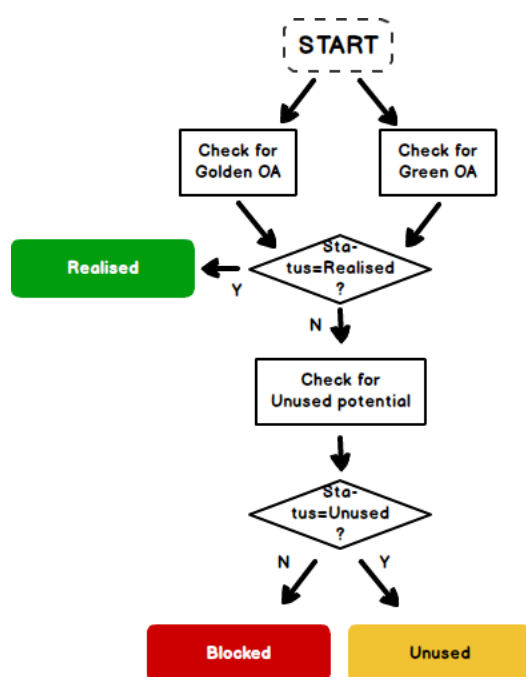
Please note, that although the diagram above indicates that validation for Golden and Green Open Access takes place in parallel, the actual implementation is, that Golden is validated before Green.

Each of the steps illustrated above are workflows of their own. They are described individually below.

4.1.1   Checking for Golden Open Access Potential

First, the journal registered in the publication metadata record is checked against DOAJ. If present, the publication is considered one with a (Golden) Open Access potential, and the potential is considered to be Realised.

To determine the type of realisation, DOAJ API is requested for the journal, and JSON response element apc{average_price} is checked. Below, this element is referred to in shorthand notation 'apc_price'.

If apc_price has a value bigger than zero, the type of realisation is considered to be **Golden with APC**. Otherwise, it is **Golden without APC**.

The associated – simple - workflow can be depicted as follows:
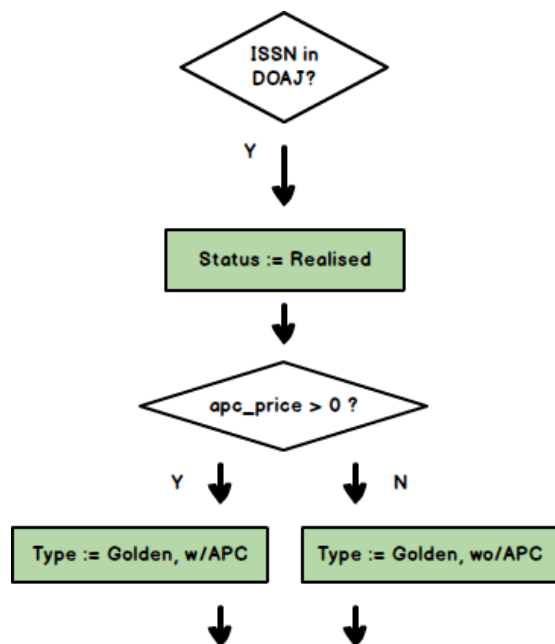
4.1.2   Checking for Green Open Access Potential

Green Open Access validation of a publication record involves inspecting the element /ddf_doc/oa_link. Below, it will be referred to with the shorthand notation //oa_link.

Records may contain zero, one or more //oa_link elements. The combined worflow for validating Green Open Access is as follows:
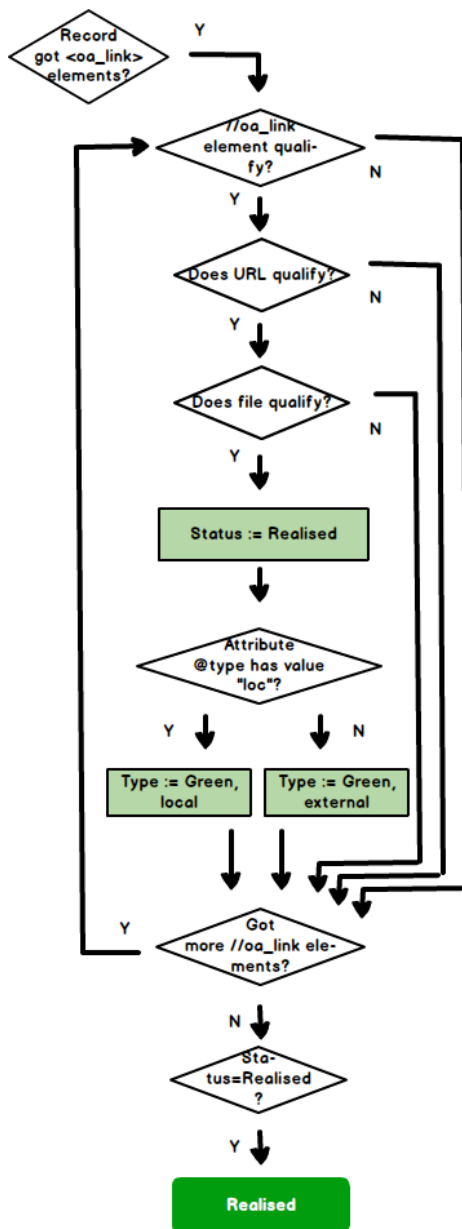
Three decisions in this workflow has to do with qualification. These three decisions are made following sub-workflows described below. For each file that pass all three decisions successfully, giving the record status Realised, the Type of (Green Open Access) realisation for this file is determined. This fourth decision is also described below.
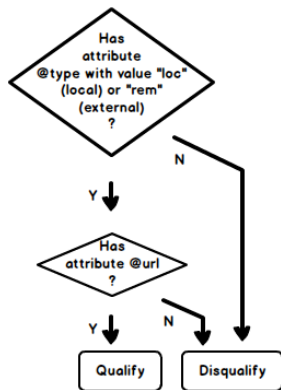
Decision 1: //oa_link element qualify?
A *qualified //oa_link element* is a //oa_link element
- with attribute @type having an acceptable value ("loc" for local or "rem" for remote" – not "doi" for DOI), and
- with a @url attribute that has a value.
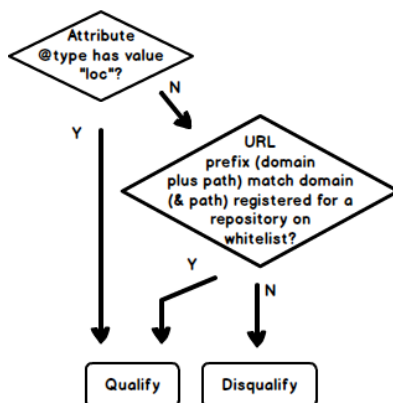Checking for qualification can be illustrated with the following workflow:

//oa_link element qualify?

Decision 2: Does URL qualify?

A *qualified URL* is either a URL to a local repository or a URL to an external repository that has a prefix (domain name and potentially also path) registered for a repository on the list of accepted external(/remote) repositories (the Whitelist). Checking for qualification can be illustrated with the following workflow:
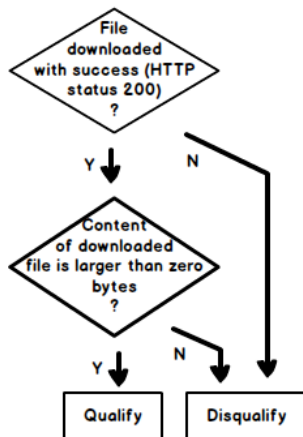
URL qualify?

Decision 3: Does File qualify?

A *qualified file* is a file that
   - can be downloaded by a computer
   - where the content of the downloaded file has size bigger than zero

Checking for qualification can be illustrated with the following workflow:
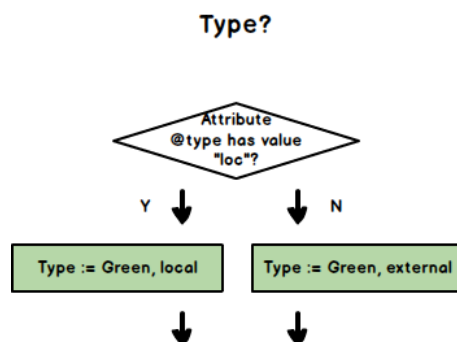
File qualify?

Decision 4: Determining the type of realisation

The type of realisation is determined by attribute //oa_link/@type:

- If this attribute has value "loc", the type is **Green Open Access from local repository**,
- otherwise it is **Green Open Access from external repository**.

Illustrated by the following workflow:

Type?

4.1.3   Checking for Unused & Blocked Potential

If the record has no Realised Open Access Potential, the record is examined to determine if the potential is Unused or Blocked.
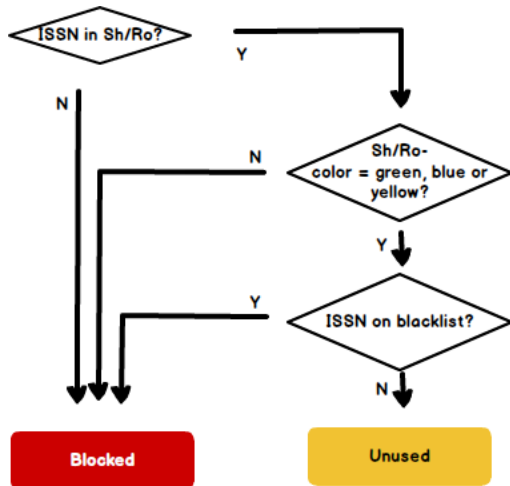
The Open Access potential of the publication is derived from the the Open Access potential of the journal registered in the publication metadata record, as registered in the Sherpa/Romea dataset (c.f. http://www.sherpa.ac.uk/romeoinfo.html).

Rules applied:

- If the ISSN of the journal is registered in Sherpa/Romeo with color code green, blue or yellow, the journal is considered one with Open Access Potential, and the publication metadata record is considered one with an **Unused** Open Access potential.
    - o   An Exception to this rule is, if the ISSN is registered on the list of accepted journals with extended embargo periods (the Blacklist). If so, the record is reclassified to Blocked
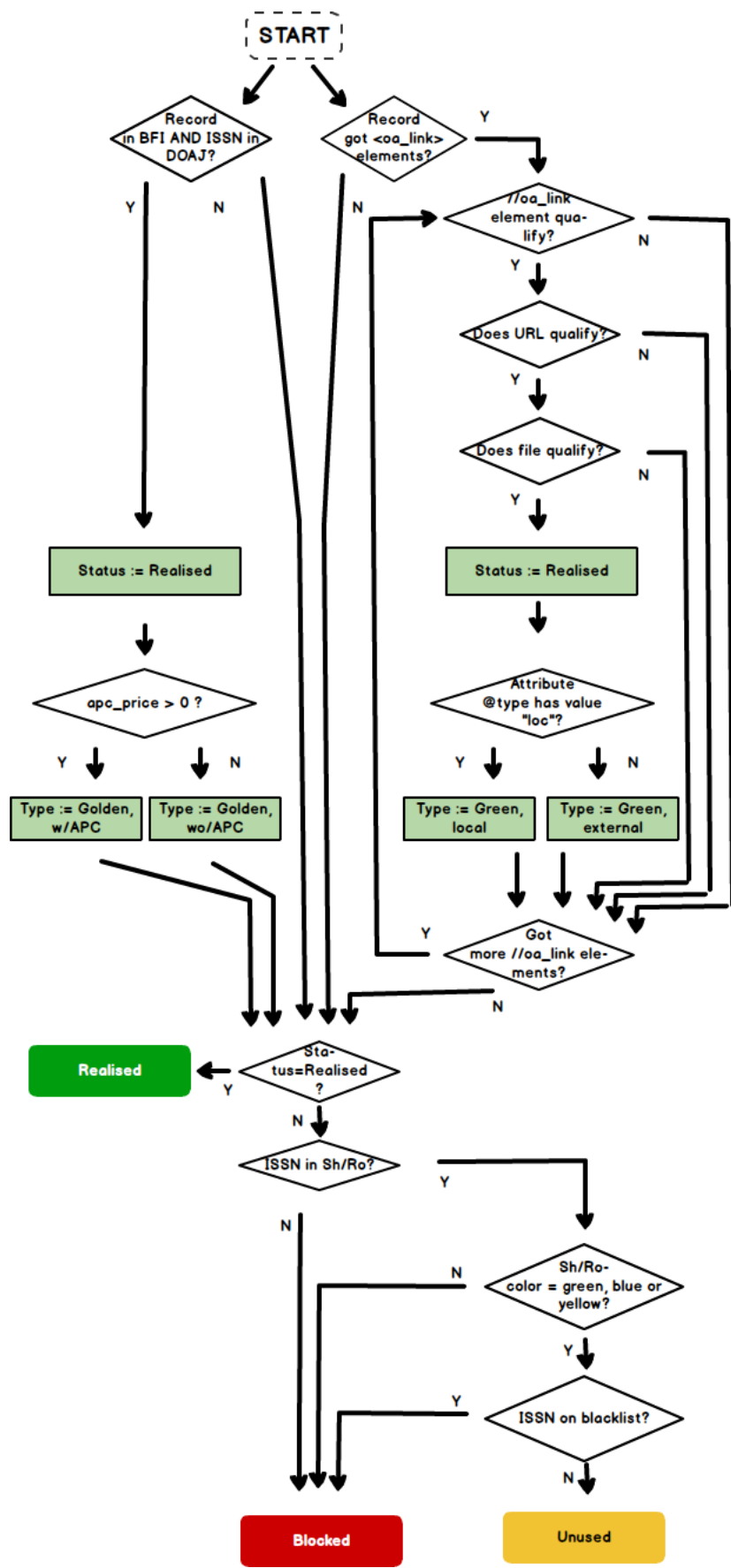
15

- If the journal is registered in Sherpa/Romeo with a different color code or not registered at all, the journal does not have a clear Open Access potential, and the publication metadata record is considered to be one with an **Blocked** Open Access potential.

This validation can be depicted as follows:


4.1.4   Checking Open Access Potential – Combined

Thus, the combined decision workflow for determining the Open Access potential of a record is:

START

Record in BFI AND ISSN in DOAJ?

Record got <oa_link> elements?

//oa_link element qualify?

Does URL qualify?

Does file qualify?

Status := Realised

Status := Realised

apc_price > 0 ?

Attribute @type has value "loc"?

Type := Golden, w/APC

Type := Golden, wo/APC

Type := Green, local

Type := Green, external

Got more //oa_link elements?

Status=Realised?

Realised

ISSN in Sh/Ro?

Sh/Ro-color = green, blue or yellow?

ISSN on blacklist?

Blocked

Unused

4.2 Open Access Classification – National and Main Research Area Level

Publication metadata records in the set of scoped records excluding duplicates correspond to clusters of one or more records from the set of scoped records including duplicates.

After classifying each of the records of the set of scoped records including duplicates according to Open Access potential and its realization, clusters inherit classifications according to a **"best-classification-wins" algorithm**, using the following decision workflow:
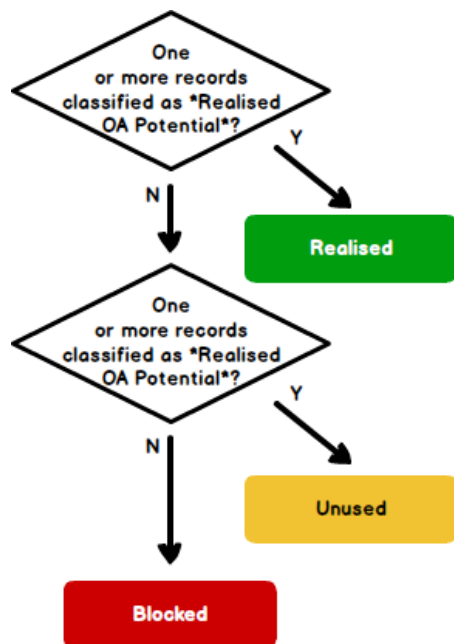
For clusters classified as Realised, the type of realisation for the cluster is also inherited from the records of the cluster.

The inheritance is done **by union**: Any type of realisation associated to any record in the cluster that are classified as Realised, are also associated with the cluster as a whole.

5 Process 4: Quality Assurance
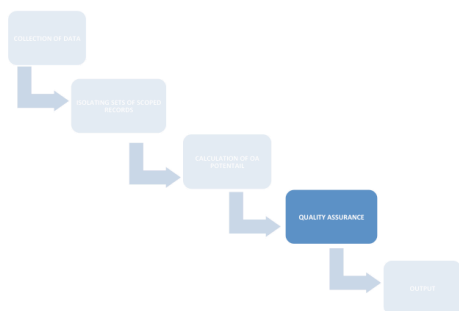
The results of the Open Access Indicator have been subjected to quality assurance. For a description, please refer to the Overview documentation

6 Process 5: Output
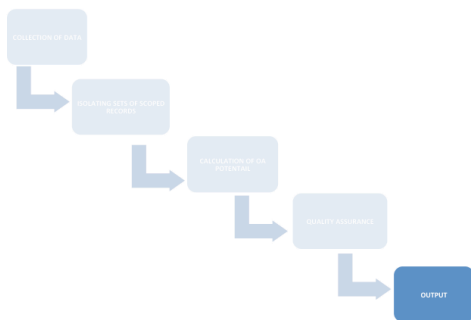
As output, the Open Access Indicator produce a number of data reports as well as web-friendly visualisations of the summations of these.

6.1 Data Reports for download

Five data reports are produced:

1) Summations:: The sets of scoped records, aggregated and distributed on <mark>Realized</mark>, <mark>Unused</mark> and <mark>Blocked</mark> Open Access potential
   a. ***Nationaly*** (set of scoped records *excluding* duplicates)
   b. Distributed on ***Main Research Area*** (set of scoped records *excluding* duplicates)
   c. Distributed on ***the universities*** (set of scoped records *including* duplicates)

2) Detailed foundation for (a) and (b): Total list of publication records in the ***set of scoped records excluding duplicates***

3) Detailed foundation for (c): Total list of publication records in the ***set of scoped records including duplicates***

4) The list of ***accepted external repositories*** (The Whitelist) used for the calculation

5) The list of ***accepted journals with extended embargoes*** (The Blacklist) used for the calculation

6.2 Web Dissemination

The summations of the Open Access Indicator are visualised on https://oaindikator.dk/, from where data reports can be downloaded as well.

7 Appendix A: The Fulltext Download Sub Process

All the fulltexts registered (by its URL) in the scoped set of publication metadata records are attempted downloaded in a single sub process.

This sub process is implemented in the following way:

- Fulltexts are downloaded one by one (serial; not in parallel)

- Fulltexts are downloaded in a "University Round Robin" fashion:
  - one fulltext from university 1
  - one fulltext from university 2,
  - one fulltext from university 3,
  - ...,
  - one fulltext from university N,
  - one fulltext from university 1,
  - one fulltext from university 2,
  - ...,
  - one fulltext from university N,
  - ...
  - ...

All downloads are done automatically by the OA Indicator download robot.

Any repository holding the fulltexts (either the research databases of the universities or external repositories) can identify a download by the OA Indicator robot by:

- IP address: 192.38.67.38