Open Access Indicator for 2014

Part 2

Technical Description of Data Foundation, Processes and Output

Revision 1 of 10 March 2016

0 Preface

The National Steering Group for Open Access[1] has proposed the Danish Agency for Science, Technology and Innovation and Denmark's Electronic Research Library, to develop a Danish Open Access Indicator. The intention is to support the implementation of the national Open Access strategy[2] - cf. the strategy's statement on monitoring: "*The implementation of Open Access is to be monitored on an ongoing basis to ensure that all parties make a maximum effort to develop and disseminate free accessibility to Danish research findings.*"

The Open Access Indicator is calculated once per year with the target field: *Scientific and peer reviewed articles and conference contributions in journals and proceedings with ISSN.*

In the context of Horizon 2020[3], EU requires that Open Access be established within at most 6 months after publication for the areas of science, technology and health and within at most 12 months for the social sciences and humanities. This delay is caused by many journals maintaining so-called embargo periods, where they exclude researchers from establishing Open Access to the articles before the end of the embargo period.

As the OA Indicator is calculated once annually for all publications within its target field, it is designed to accept a one-year delay in Open Access to the publications. Consequently, the OA Indicator for 2014 is calculated early January 2016 in order to accommodate a full year embargo period also for publications from December 2014. In practice this means that publications from January 2014 could have embargo periods all the way up to 24 months and still be credited by the OA Indicator.

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: Below, the notion of the indicator's "target field" is expressed using the term "set of scoped records".

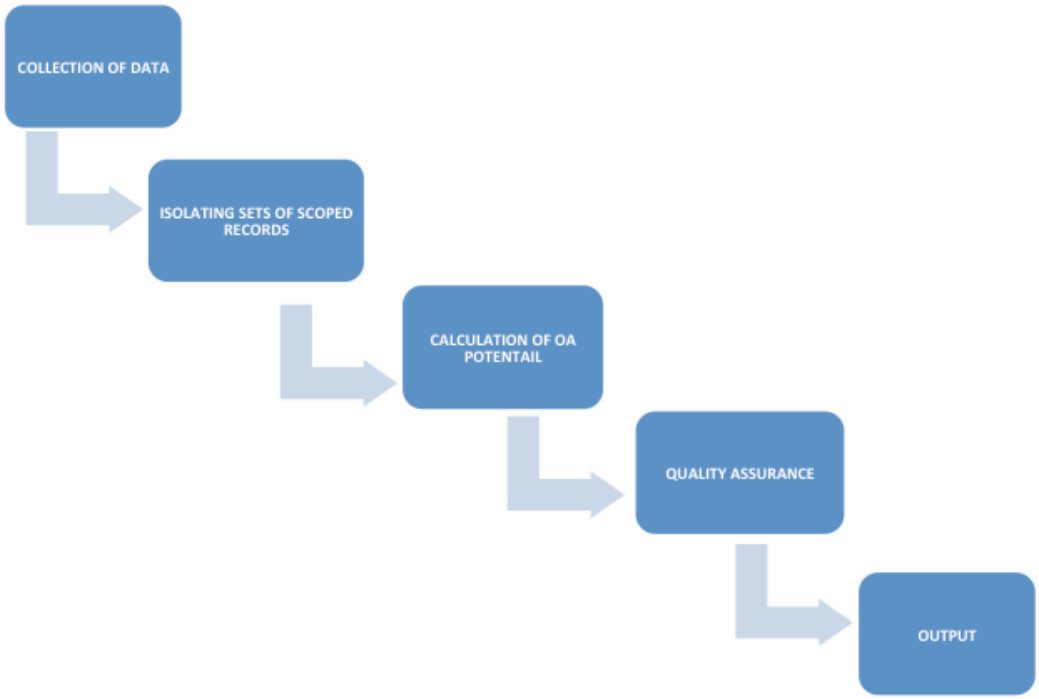
Queries regarding the indicator may be directed to

Jonas Bak/Hanne-Louise Kirkegaard
6th Division: Research Policy
Danish Agency for Science, Technology and Innovation
Bredgade 40
DK-1260 København K
Email: jonb@fi.dk/ hki@fi.dk

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

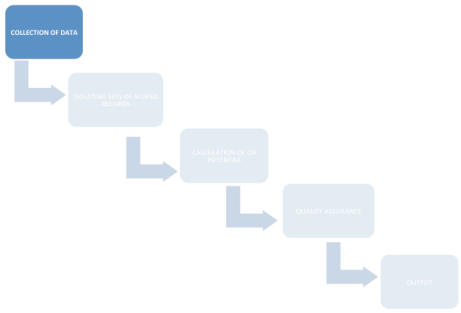
1 Main Processes

The activities of the OA Indicator can be broken down into these five main processes.



The five main processes are described in further detail on the sections below.

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 four national and international sources. The data foundation are 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.

2.1.1 Requirements on Universities – Data 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 exists 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 2014:

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 are used by the OA Indicator as an authorative list of Golden Open Access Journals. 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/)

2.2.2 Sherpa/Romeo (Sh/Ro)

Sh/Ro are 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 three 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 Area's registered in the metadata for the publications
- To ensure that articles published in DOAJ-validated journals can be considered scientific and peer-reviewed (BFI-level 1 or 2).

Parameters of the data collection:

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

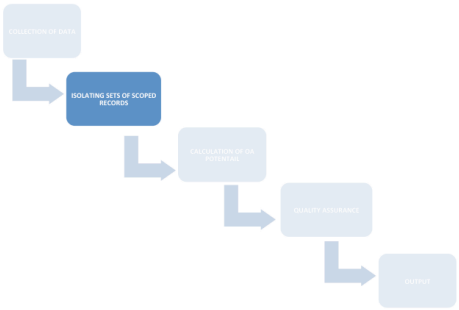
2.3 This Years Complete Data Collection

Summary of the data collection for the OA Indicator for 2014:

Source	Protocol	Ver.	Format	Ver.	Collection Date	Records
AAU	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	7019*
AU	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	13223*
CBS	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	2386*
DTU	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	7057*
ITU	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	351*
KU	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	13214*
RUC	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	1751*
SDU	OAI-PMH	2.0	DDF-MXD	1.3.0	8/1 – 2016	7268*
DOAJ	OAI-PMH	2.0	DC	%	8/1 – 2016	10989
Sh/Ro	HTTP	%	Proprietary	%	8/1 – 2016	25076
BFI	HTTPS	%	Proprietary	%	9/10 - 2015	27079

* With Submission Year 2014

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 are 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” eller “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 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 proces. The ambition of this proces 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 contain 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, titel, undertitel, ISSN, publikationsår, 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 differs. 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 dedup 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 are 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 have 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, The MRA of the cluster is chosen by random among those MRAs.

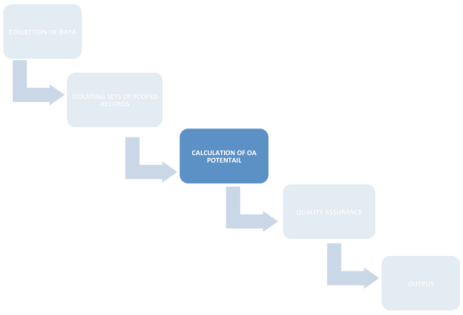
This algorithm ensures, that the OA Indicator solves potential MRA-conflicts respecting to the largest extend possible the corresponding MRA-conflict resolutions done by BFI.

3.3 This Years Sets of Scoped Records

Dataset	Records
Total number of publication records collected from the universities	52.269
Set of scoped records including duplicates	24.362
Set of scoped records excluding duplicates	21.943

For further details, see section on Data reports.

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 – are 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 are classified according to how the publication realise its Open Access potential.

There are three values for this classifications, and they are colorcoded using green, yellow and red (traffic light):

- **Realised** Open Access potential
- **Unused** Open Access potential, and
- **Unclear** Open Access potential

4.1 Open Access Classification – University Level

For any record in the set of scoped records including duplicates, the Open Access potential is initially established by validating potential Golden Open Access and only subsequently validating Green Open Access.

4.1.1 Golden Open Access Validation

First, the journal registered in the publication metadata record are checked against DOAJ. If present, and if the publication record achived a level 1 or level 2 BFI classification, the publication is considered one with a (Golden) Open Access potential, and the potential is considered to be **Realized**.

If not, the record is examined for potential Green Open Access and its potential realization.

4.1.2 Green Open Access Validation

Initially, the record is checked for ***OA qualified links,*** which may point to ***OA qualified files.***

The presence of ***OA qualified links*** are determined by the following rule:

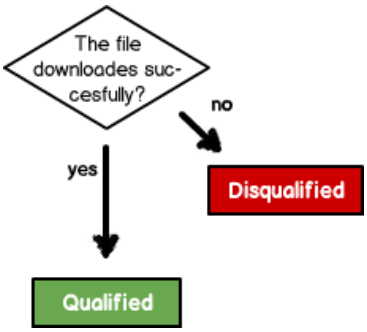
1) Are link(s) to fulltext file(s) registered in the publication metadata record?
    a. ***Rule applied***: One or more elements /ddf_doc/publication/digital_object/uri exists in the publication metadata record
2) Can registered links be considered links that can demonstrate Open Access?
    a. ***Rule Applied:*** Any link registered in elements /ddf_doc/publication/digital_object/uri are accepted.

Please note: (1) implies that only links pointing to files deposited in the universities research databases qualify. Links to files deposited into external (subject specific) repositories do not qualify. (This will change in the OA Indicator for 2015).

The presence of ***OA qualified files*** are determined by the following rule:

1) The file(s) pointed to by OA qualified links can succesfully be downloaded
    a. ***Rule applied***: The file(s) can be downloaded by computer by following the link.

Thus, the decission workflow for determining the presence of ***OA qualified files*** are simple:



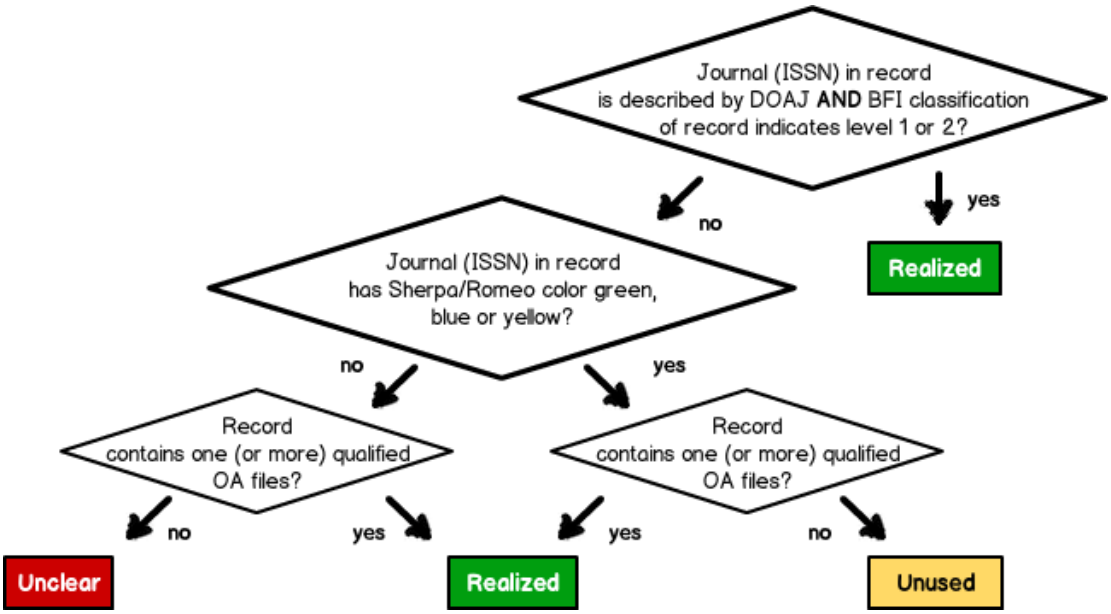
Based on this, the (Green) Open Access potential of the publication metadata record are determined according to the following procedure:

- If the record contains one or more OA qualified links pointing to OA qualified files, the publication are considered to be one with a **Realized** Open Access potential.

- Otherwise, the Open Access potential of the publication are 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).

    - ***Rule applied:***
      If the journals ISSN are 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 are considered one with an **Unused** Open Access potential.
      If the journal is registered 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 are considered to be one with an **Unclear** Open Access potential.

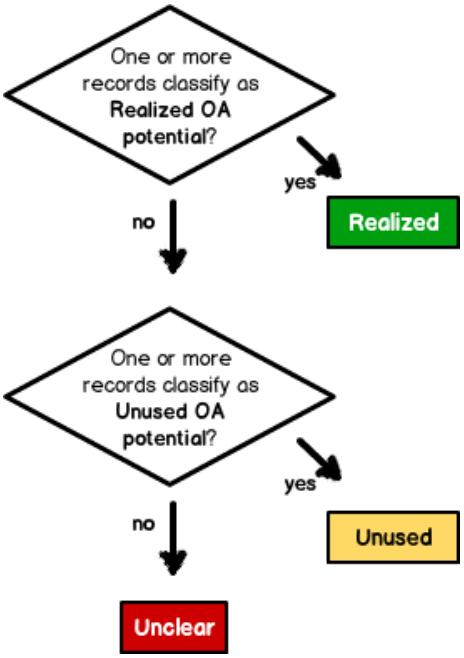
Thus, the combined decission workflow for determining Green Open Access potential are:



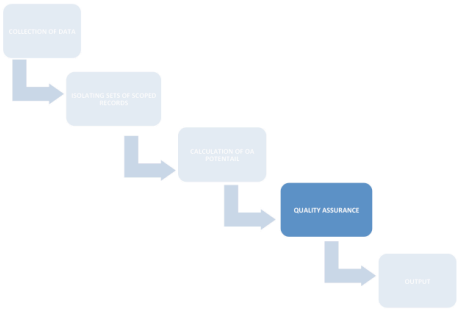
4.2 Open Access Classification – National and Main Research Area Level

Publication metadata records in the set of scoped records excluding duplicates corresponds 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 decission workflow:



5 Process 4: Quality Assurance

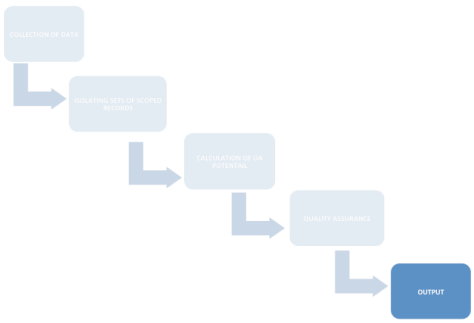


The results of the Open Access Indicator have been subjected to the following quality assurance meassures:

- **Data Foundation**. The collected data and the registered links and their resolvability back to the universities research databases, has been tested. The tests have been based on sampling across the universities.

- **Downloaded fulltext files**. A selection of the downloaded fulltext files have been inspected to ensure that they can indeed be considered files representing the scientific article – in a complete and readable fashion. The test have focused on files that, based on simple computerbased analysis, could seem to deviate suspiciously from the metadata registered for the publication (page numbers, file sizes, etc).

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.

The Danish Research Database (http://forskningsdatabasen.dk/) are used as dissemination platform for the visualisations and the reports.

6.1 Data Reports for download

Three data reports are produced:

1) Summations:: The sets of scoped records, aggregated and distributed on **Realized**, **Unused** and **Unclear** 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 ©: Total list of publication records in the ***set of scoped records including duplicates***

6.2 Web Dissemination via The Danish Research Database

The summations of the Open Access Indicator are visualised on http://open_access.ddf.dtic.dk/en/open_access/overview, from where data reports can be downloaded as well.