

Annex 2 – Context and Architecture

1 Introduction

1.1 KB activities related to this tender

The primary collection of the KB consists of publications from or about the Netherlands. Since a couple of decades such publications are increasingly and preferably collected in digital form.

The KB has implemented processes & workflows to acquire these so called *born digital* publications (like e-journals and e-books) directly from publishers and institutions (e.g. government publications). Digital research publications and other research output are harvested from the institutional repositories of Dutch universities and research institutes.

Additionally, the KB harvests a selection of websites relevant to Dutch heritage. Plans are being developed to extend this harvesting to comprise large parts of the Dutch .nl domain, as a depot for the Dutch web. Pilot projects for harvesting (relevant) Dutch social media content are being considered.

The KB works with international scientific publishers and other partners (cf. thekeepers.org) to safeguard the long-term access to journal content through its e-Depot service.

The KB keeps investing heavily in digitizing its and other institutions printed collections, through government funded national digitization programs, like *Metamorfoze* (<https://www.metamorfoze.nl/english>), or in collaboration with other partners (e.g. Google Books). The digitized collections contain hundreds of thousands of books and journals, and millions of newspaper articles and issues. These collections will keep growing substantially in the coming years.

Currently the ensemble of the KB's digital collections are managed and stored using a large variety of storage systems and media, and through different workflows and processes.

Through its *national bibliography*, the KB offers an authoritative register of all publications from or about the Netherlands, in printed and/or digital form.

The ensemble of its digital collections are available to users through various access facilities. The public web portal *Delpher* (www.delpher.nl) offers access to the integral texts of Dutch books, journals and newspapers.

The *Digital Library for Dutch Literature* (<http://www.dbnl.org>) is a website about Dutch language and Dutch literature. This website offers thousands of literary texts, secondary literature and additional information, like biographies, portrayals etc.

Data services make the total corpus of the KB digital collections available for (academic) research. The KB will partner with research institutes and other organizations (e.g. in CLARIAH www.clariah.nl) to integrate its data services in national and European *research and heritage infrastructures*.

Table 1 and 2 show estimates for the total volume of KB's *born digital* and *digitized* collections, currently and in the 2025 forecast.

Collected born digital items by type	Current volume in items	Estimated volume in 2025
E-books (titles)	33 395 books	100 000 books
E-Newspapers (titles)	-	~ 50 newspapers
E-Journals (titles)	11 755 titles	20 000 titles
E-Journals (articles)	28 242 959 articles	50 million articles
Dutch research output	1 095 751 files	2 500 000 files
TBs e-books, e-journals	44,26TB	90 TB
Number of files e-books, e-journals	488 272 000 files	1 billion files
Average growth e-books, e-journals (TB)	16%	
Average growth e-books, e-journals (files)	20%	
Websites (periodic harvest)	14 983 sites	6 million sites
Number of files websites	2 million files	6 million files
TBs websites	36 TB	108 TB
Average growth websites	20%	
Total storage volume (in TB)	80,26 TB	200 TB

Publications by type	Current volume in items	Estimated volume in 2025
<i>Digitized</i>		
Books	10 million pages	20 million pages
Newspapers	20 million pages	30 million pages
Magazines	6 million pages	10 million pages
Written manuscripts	200 000 pages	500 000 pages
Parliamentary papers	2,4 million pages	2,4 million pages
News bulletins	1,8 million items	1,8 million items
Total number of files	496 945 715 files	1.5 billion files
Average growth digitized material	20%	
Total storage volume (in TB)	500 TB	1500 TB

1.2 KB Ambition for the next 7-10 years

KB's digital collections are considered a major part of Dutch national heritage, and as such need to be safeguarded for long-term availability.

The KB aims to improve the management and (long-term) preservation of its entire digital collections. KB will use the OAIS reference model (ISO 14721:2012) as a guideline to provide the concepts, organizational models and systems for this. Quality assurance for this improvement will involve external auditing and certification as a trusted digital repository of its digital collections.

To help with this task the KB seeks to implement a *Digital Asset Management and Preservation System* (DAMPS). The system will be based on the OAIS reference model and principles, and will provide the necessary tools to help with the management and long-term preservation of its entire digital collections of *born digital* and *digitized* publications and other assets.

Figure 1 shows an abstract view of the environment and functions of the envisaged DAMP system, in terms of the OAIS reference model.

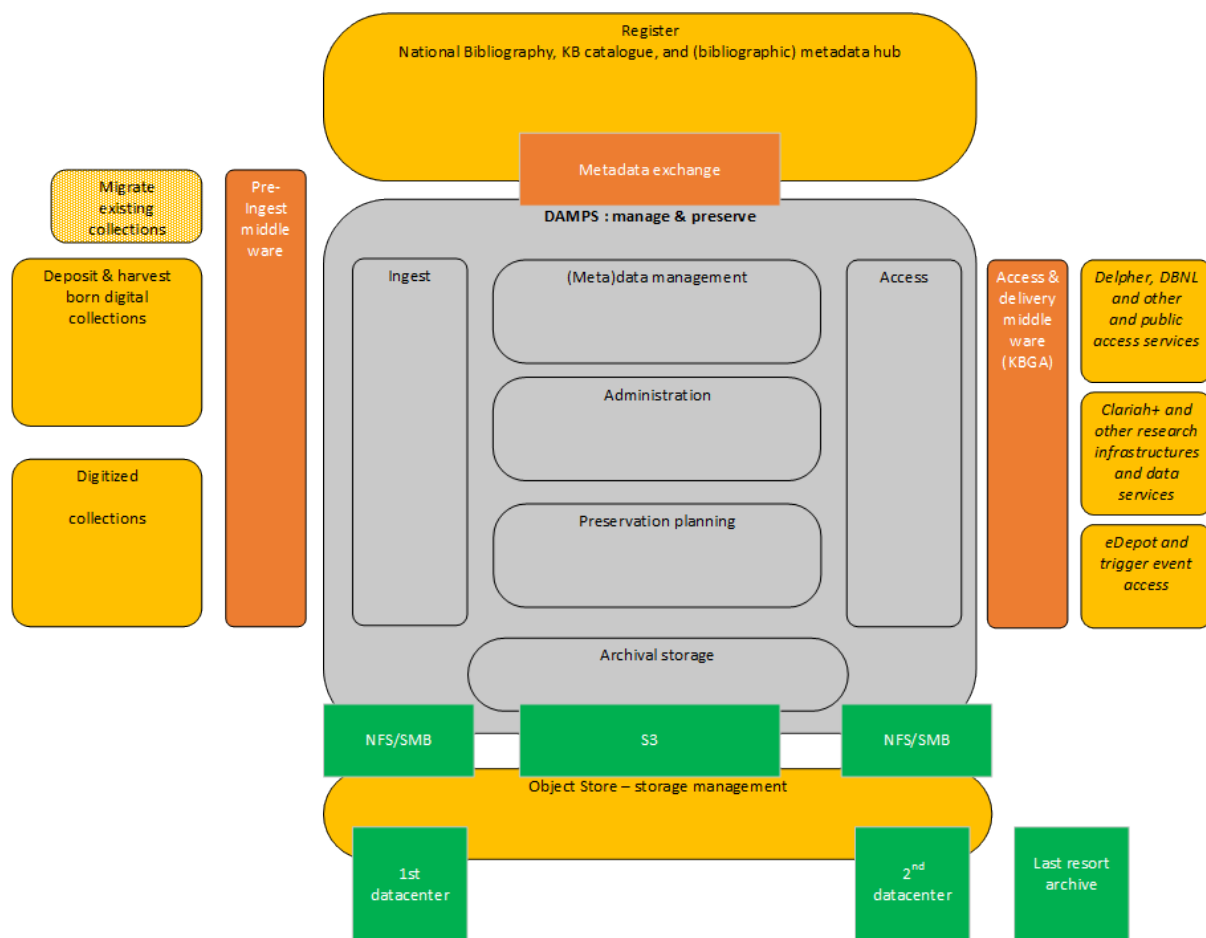


Figure 1: abstract view of the environment and functions of the envisaged DAMP system, in terms of the OAIS reference model

For the next 7-10 years the main priority is to reduce the complexities of the current systems and bring all digital collections under administration.

Access will for the time being remain through the existing systems for continuation of current services.

Preserving digital heritage collections for long-term availability requires constant and active curation of these collections for a period that exceeds by far the expected lifetime of any actual DAMP system or supplier thereof. Any potential supplier should keep that in mind. The KB will use different archival storage or systems for management and preservation of its collections in the future.

1.3 Interfaces

1.3.1 Ingest

We expect the DAMPS to offer *Ingest* functionality, directly interfacing with *Deposit* workflows & processes from KB's suppliers (publishers & institutions), KB's *Harvesting* workflows & processes for web- and repository harvesting, and KB's *Digitizing* workflows & processes.

Additionally we expect the DAMPS *Ingest* functions to offer adequate tooling for migrating KB's existing digital collections into the archival storage managed by the DAMPS and the associated *object store*, using automated *Ingest* workflows.

1.3.2 Archival Storage

We expect the DAMPS *Archival Storage* function to interface with the storage archive using the S3 protocol¹. The storage system itself will manage data copies on several locations and a data disaster recovery system.

1.3.3 Metadata Exchange

We expect the DAMPS to synchronize the (bibliographical) metadata associated with the information objects in its digital collections with KB's metadata hub, and cataloguing workflow & processes.

Currently KB is using OCLC's *Gemeenschappelijk Geautomatiseerd Catalogiseersysteem* (union catalogue) as its bibliographic metadata hub and for its cataloguing workflow & processes. KB is preparing migration to a successor system in the foreseeable future (expected in 2021).

1.3.4 Access

DAMPS will have an interface with the KBGA (KB Gegevens Architectuur). The KBGA is the set of middleware systems that is used by the KB services² to offer clients access to KB's digital collections. DAMPS should feed the KBGA with the necessary information for this.

The KBGA consists of a number of functional components:

- **MDO** (MetaData Opslag): storage of XML records with the derived structural and descriptive metadata of the material to be provided. MDO may be queried by an OAI-PMH interface. At present, KBMDO can be fed by serving the XML records, packaged in a simple XML wrapper, to an http interface (POST) or by placing them in a load directory.
- **Solr/jSRU**: Solr provides search indexes for full-text content and the descriptive metadata of the material to be made available. The search indexes are populated by harvesting metadata from MDO via OAI-PMH. The relevant OCR text files are collected for indexing on the basis of the structural metadata. The Solr indexes can be queried by the KB services via an SRU interface (jSRU is the name of KB's implementation of the SRU protocol).

¹ https://en.wikipedia.org/wiki/Amazon_S3

² This includes, for instance, a web portal such as Delpher, but also research data services for text and data mining for researchers (TDM).

- **Resolver:** Access images and texts can be retrieved for users over http/https. The Resolver ensures that the objects on the web storage can be accessed with the help of persistent identifiers, independently of the actual file locations. The Resolver acts here as a proxy.
- **Web storage:** the storage environment for access files. A service (called the ABCstore) is available for storing files to be provided on the web storage by the Resolver which ensures that the files offered are placed on a network share with sufficient storage space available. This service also ensures that the Resolver can connect the persistent identifier of the file to the physical location.
- **Image Viewer:** The Image Viewer provides an API, used by end user services such as Delpher, for scaling and cropping access images. Therefore, the Image Viewer will also be used to convert the access images available in the storage in JPEG2000 format into the more common jpeg format.
- **SeSAM:** SeSAM is an underlying system that enforces whether or not access is granted under Rule Based Access Control (RBAC) to the available material through both the Resolver and MDO on the basis of the context, metadata and policies.

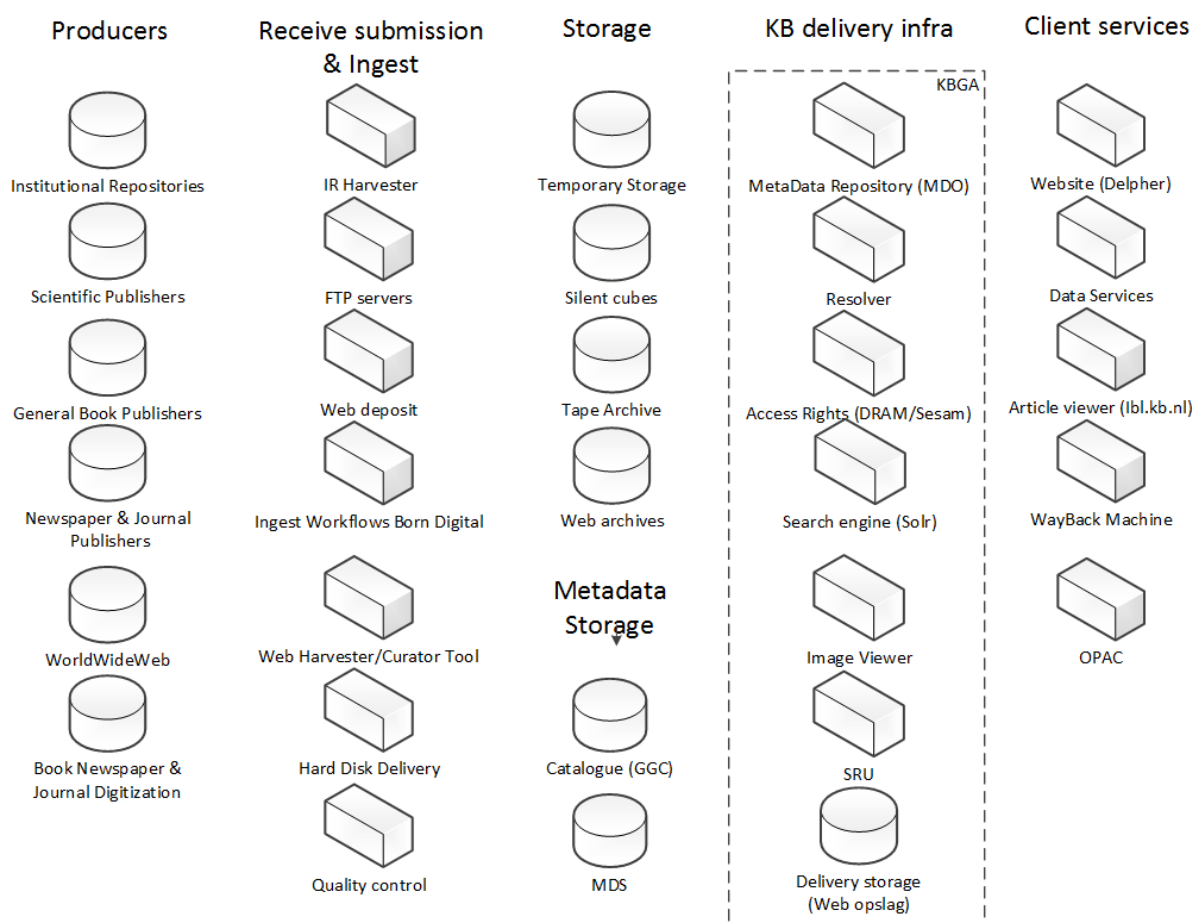
KGBA is used very intensively. On a normal day, services such as Delpher attract between 3000 and 4000 unique visitors which generate many millions of requests per month with the current set-up: Resolver 250 million, Image Viewer 70 million, jSRU 35 million. MDO (OAI-PMH) 60 million, SeSAM 400 million.

Clients mainly use the Resolver and Image Viewer (for access to images) and jSRU (for searching and retrieving) through end-user services such as Delpher. This means that SeSAM is also intensively used indirectly as an underlying system.

2 Current state of digital asset management and preservation of digital collections

2.1 General

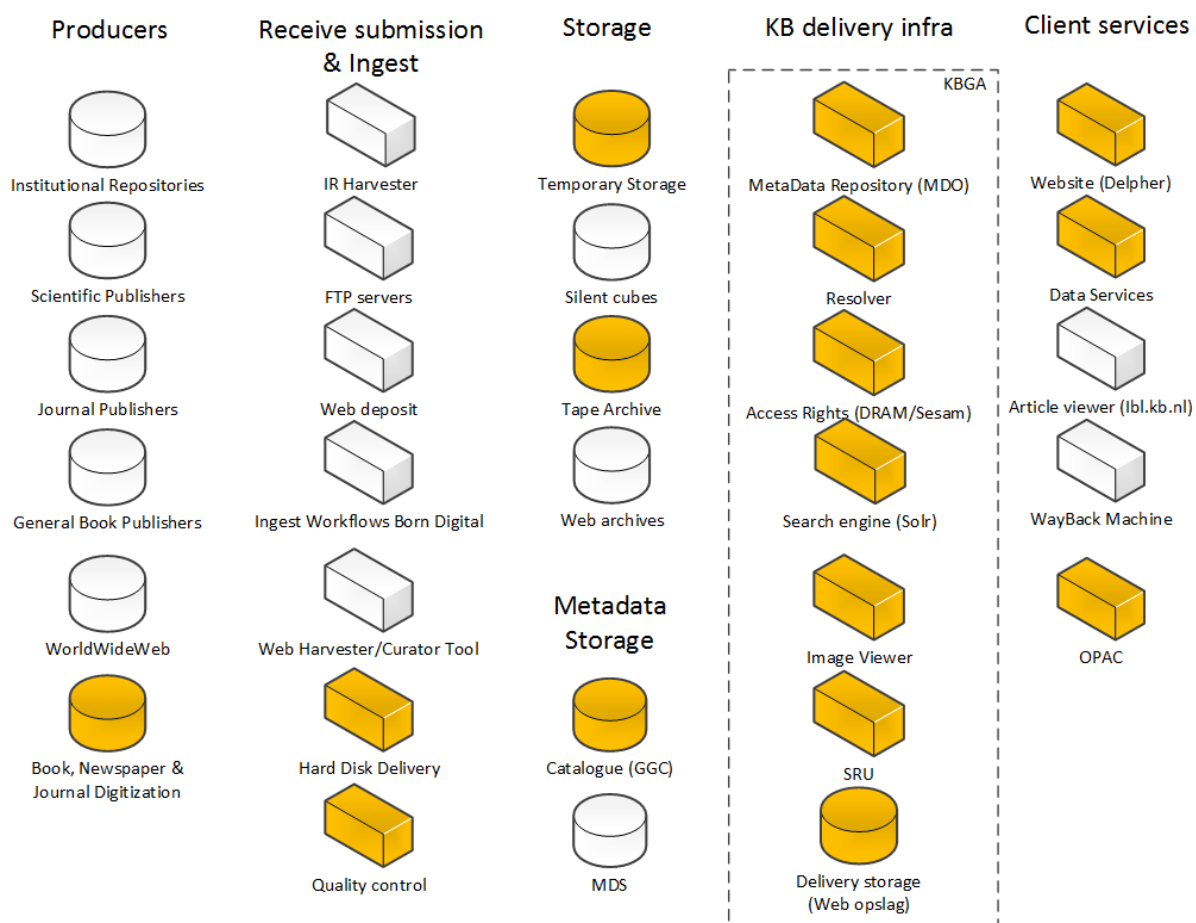
Since there are currently multiple digital collections and various work processes, the most important ones have been outlined below. The following schematic diagram should therefore serve as a framework:



2.2 Digitisation

The KB has developed initiatives over the last fifteen years for safeguarding materials that suffer from deterioration over time. This initially involved microfilms of printed materials. The introduction of new technologies at the beginning of this century has led to the replacement of the microfilms by digitisation. KB has created its own infrastructure for storing and making these materials available.

The overview below shows the most important components involved in content digitisation.



2.2.1 Digitisation and delivery (e.g. Metamorfoze)

Over the years, various projects have been carried out, each having their own process and specifications. A large part of the digitisation is focused on printed books, newspapers and journal materials.

The digitisation process produces the following type of files (for a more detailed description, see Annex 8 Data Model):

- Access images - low resolution images for each page in JPEG2000 for access purposes
- ALTO XML - technical metadata file for each page in XML format with the layout of the page concerned in accordance with the ALTO standard.
- Master images - high resolution images of each page (TIFF or JPEG2000)
- PDF file - PDF of each object, in accordance with PDF version 1.5.
- Text file - a file in a simple XML format, with information on each page (books and journals) or on each article (newspapers).
- Concordance table - CSV file with the cardinality between the files in a delivered batch.
- DMD file - XML file with descriptive metadata for each object. MPEG21 file - XML file with bibliographical and structural metadata for each object.

- METS file - XML file consisting of METS, MODS, MIX and PREMIS standards which includes bibliographical, structural, technical, rights and provenance metadata. MIX file - XML file with information about the scanner or camera used for creating the Access and Master images.
- Target images - control files in both JPEG2000 and TIFF formats

The material is checked using a control program configured by KB. Once approved, it is saved to tape storage using the ContentTool. The ContentTool is an application written by KB for communicating with the tape storage and for keeping up to date which collections have been saved.

2.2.2 Digitisation by Google

In addition to the standard digitisation process, KB has also entered into a partnership with Google to accelerate the digitisation of its collection.

The Google project is still ongoing and will lead to hundreds of thousands of extra digital works (mainly books). The digitisation is being carried out by Google. This digitisation results in other types of files than those from the standard process described above:

- hOCR - HTML file with structural metadata for each page.
- METS XML - essentially the same idea as standard XML, but with other specifications that are not present at KB.
- Text file - plain text with the text for each page.
- Image files - scans in JPEG2000 format.

2.2.3 Digital Library of Dutch Literature (DBNL)

An organisational merger in 2015 meant that the Digital Library of Dutch Literature has also become part of the KB collection. Materials that are digitised for this purpose (often at high quality) also need to be stored. In addition to the publications, there is also a lot of contextually edited information about the publications. These are also saved in various other types of files than in the standard digitalisation workflow:

- TEI XML - a file with bibliographical, structural, layout and enhanced metadata for each publication
- PDF or TIFF file for each page
- PDF files for each publication
- EPUB file for each publication
- GIF and JPEG files for access

2.2.4 Other digitisation (including Metamorfoze Retro)

KB also holds various other materials in addition to its large collections of books, newspapers and journals. These include manuscripts and maps, but also cassette tapes and CDs which are bundled as a supplement to printed works. KB also digitises these to combat deterioration over time. The formats and the associated metadata vary according to the type of material, and are sometimes extremely minimal. These are relatively small in number compared to the large collections of books, newspapers and journals.

2.3 Born digital

Since material is increasingly only being published in digital format, KB has also designed processes to acquire and store this material too.

KB has been working on preserving born digital scientific publications since the inception of the first “e-Depot” (DIAS, Digital Information and Archival System) in 2003. It has also done work since 2010 on modernisation and improvement, which has resulted in the current “Digital Warehouse 1.5” (DM 1.5).

Born Digital material is processed in DM 1.5 from:

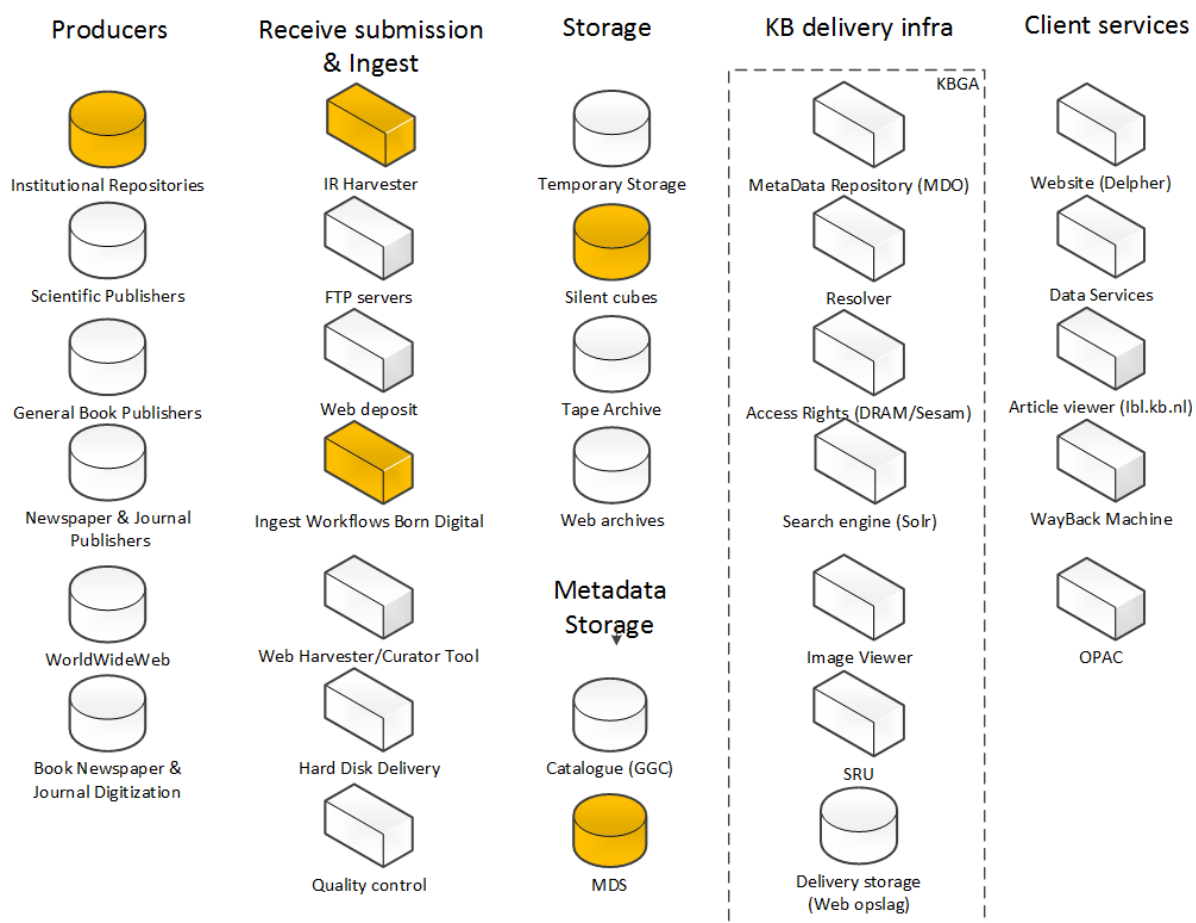
- Institutional repositories
- Publishers of scientific articles from e-journals or digitised journals
- Publishers of e-books
- Publishers of issues of non-scientific e-journals

This material is currently being stored on archival storage in the form of ‘Silent Cubes’ and an Oracle database with metadata on the publications (MDS).

The current set-up has a link to the union catalogue (GGC) for this material. Bibliographical metadata from the GGC is then included in the AIP.

The current solution is also limited in terms of file formats. PDF and EPUB are the preferred formats. Technical metadata is recorded from these file formats. For other file formats, technical metadata is not recorded. In the future, we must be able to record technical metadata for all types of file formats (see the requirements).

2.3.1 Institutional repositories



For long-term storage and permanent access to digital objects from the Institutional Repositories (IRs) of Dutch universities and scientific institutions, a link has been created between these IRs and the KB's e-Depot. The publications designated for inclusion in the e-Depot are offered in sets by the Institutional Repositories.

Available publications are harvested using a harvesting tool (IR Harvester) developed by KB. This tool uses OAI-PMH to first retrieve the available metadata from the GMH (Joint Harvester - a joint initiative between KB and DANS). The objects described in the metadata can then be downloaded using URLs in the metadata.

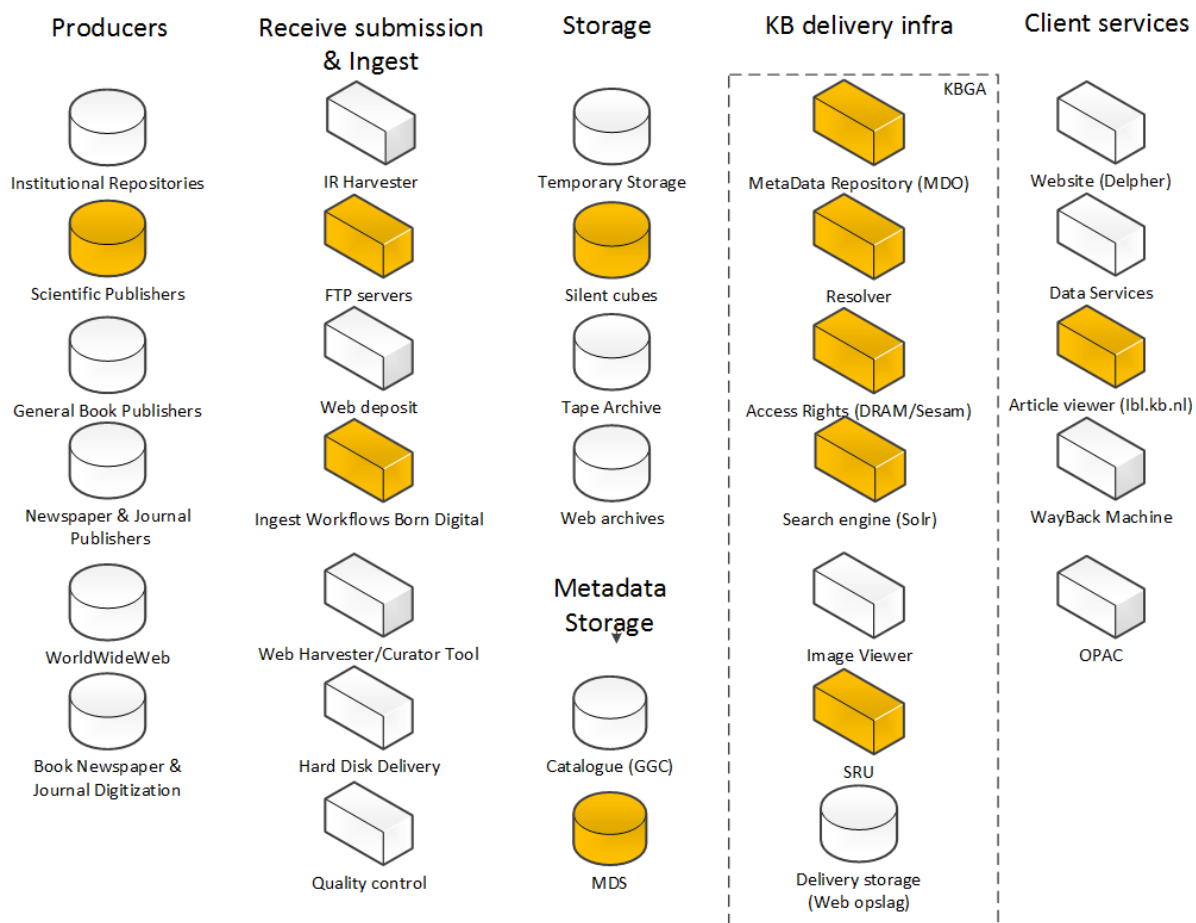
The harvester delivers an Information Package (IP) that includes the GMH metadata, as well as metadata with information on the harvesting process and the objects. Checks are carried out during the pre-ingest process based on arrangements with the repositories, such as:

- Have any files been retrieved at all?
- Is the publication Open Access (applies to all files that belong to a publication)?
- Is the object that has been retrieved the expected object?
- Has the publication already been delivered in full with the same checksums?
- Have any 0 byte files been delivered?

The publication is generally delivered in PDF format. It is also possible that extra files may be delivered in all kinds of formats as supplements.

The metadata can be delivered from MDS via OAI-PMH.

2.3.2 E-journals (academic articles)



E-journal articles by academic publishers are currently delivered through SFTP (push and pull). The form of delivery varies according to the publisher and is very diverse. For instance, one publisher supplies each delivery in a zip file and another publisher uses a folder structure (possibly with multiple journals/editions) in a zip file, or a folder structure in a TAR archive. Deliveries are also provided in non-zipped folder structures.

In all cases, the article is delivered in PDF format. Furthermore, in some cases the article is also delivered as full-text XML or as “plain text”.

A delivery may also contain attachments in various file formats.

Full-Text metadata is usually related to specific files that are additionally delivered.

Metadata can be delivered in a great variety of ways. Relevant information can be packaged in a folder structure in various XML records at different levels in the structure.

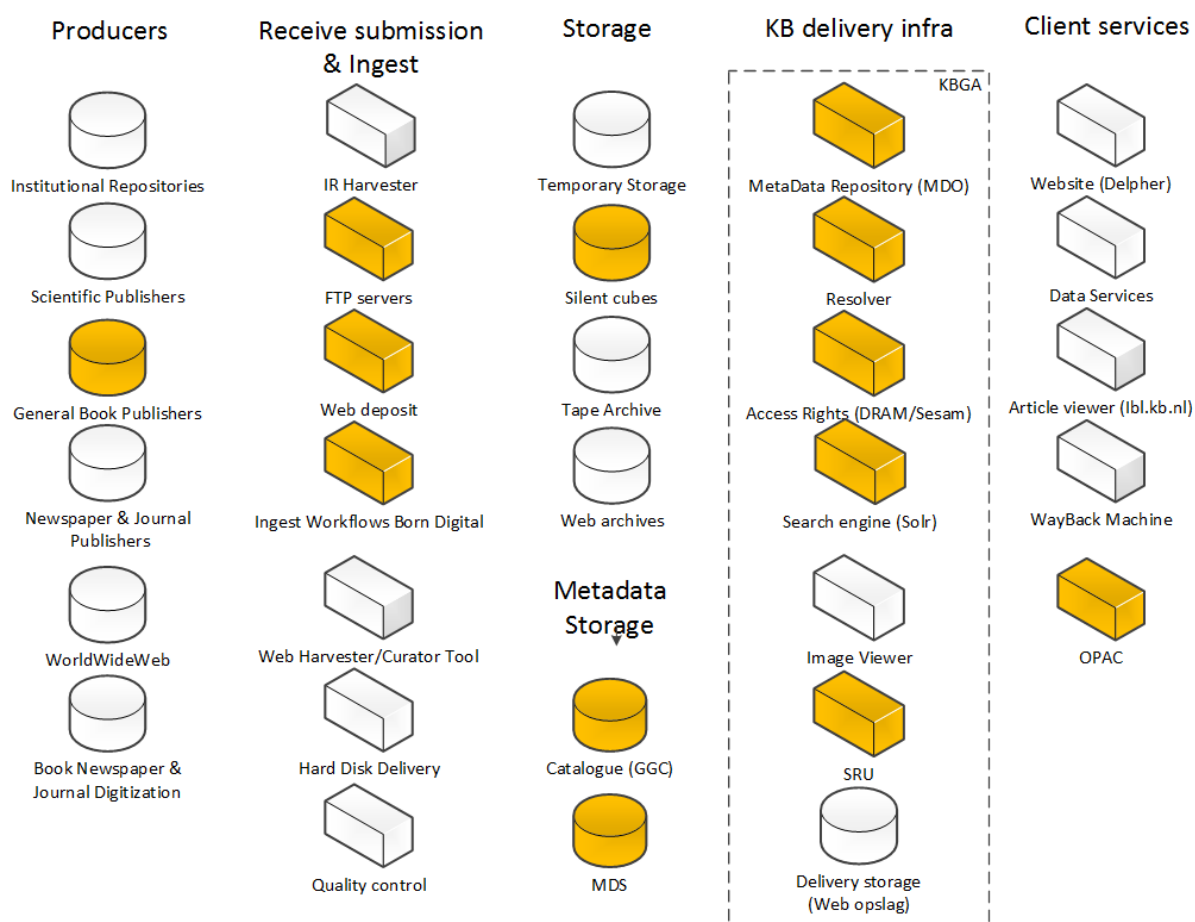
The format of the metadata varies for each supplier. For instance, one supplier uses a proprietary format and another follows a standard (NLM) format.

The metadata supplied is transformed into a format that is used in the local solution for availability (MDO). The metadata indexed by Solr is used for availability by means of a browser tool and a search interface. This enables hierarchical searches between the Journal title, Volume, Issue and Article, or in reverse order and searches for metadata.

Specific checks have been preset for each publisher, such as:

- Zip or TAR delivery: is the file not corrupt?
- Is metadata available for each article delivered?
- Is it not a 0 byte PDF?

2.3.3 E-books



The “flow” of e-books is currently made up of two types of deliveries:

1. Delivery of e-books by CB (formerly Centraal Boekhuis) from publishers affiliated with CB
2. Delivery of e-books via the Monographies Digital Desk (Webloket Monografieën) by publishers who are not affiliated with CB called Web deposit in the drawing.

1. CB flow

The CB flow is a supply of e-books from Dutch publishers according to agreements that have been made with the Dutch Publishers Association, in which CB acts as an aggregator. Metadata (ONIX) and e-books are delivered separately.

- 1x per day: The supply of metadata (ONIX) – a merger of all new and updated ONIX records (ONIX Message in a zip file);
- Possibly on hourly basis: The delivery of e-books in ePub or PDF format.

KB retrieves both of these from the FTP server at CB.

The ONIX Message is cut up into separate ONIX records that relate to one e-book and these are stored (existing record only provided with an update, where applicable).

An IP (Information Package) of an e-book and the accompanying metadata needs to be merged in pre-ingest. The metadata must meet certain conditions at that point before it can be processed. The most important condition is that an ONIX record has been checked by Bureau ISBN.

Only e-books are processed that:

- have access to ONIX (meet the conditions and must be well-formed and valid XML)
- have not already been delivered with the same checksum
- do not have a size of 0 bytes

2. Monographies Digital Desk (Webloket Monografieën)

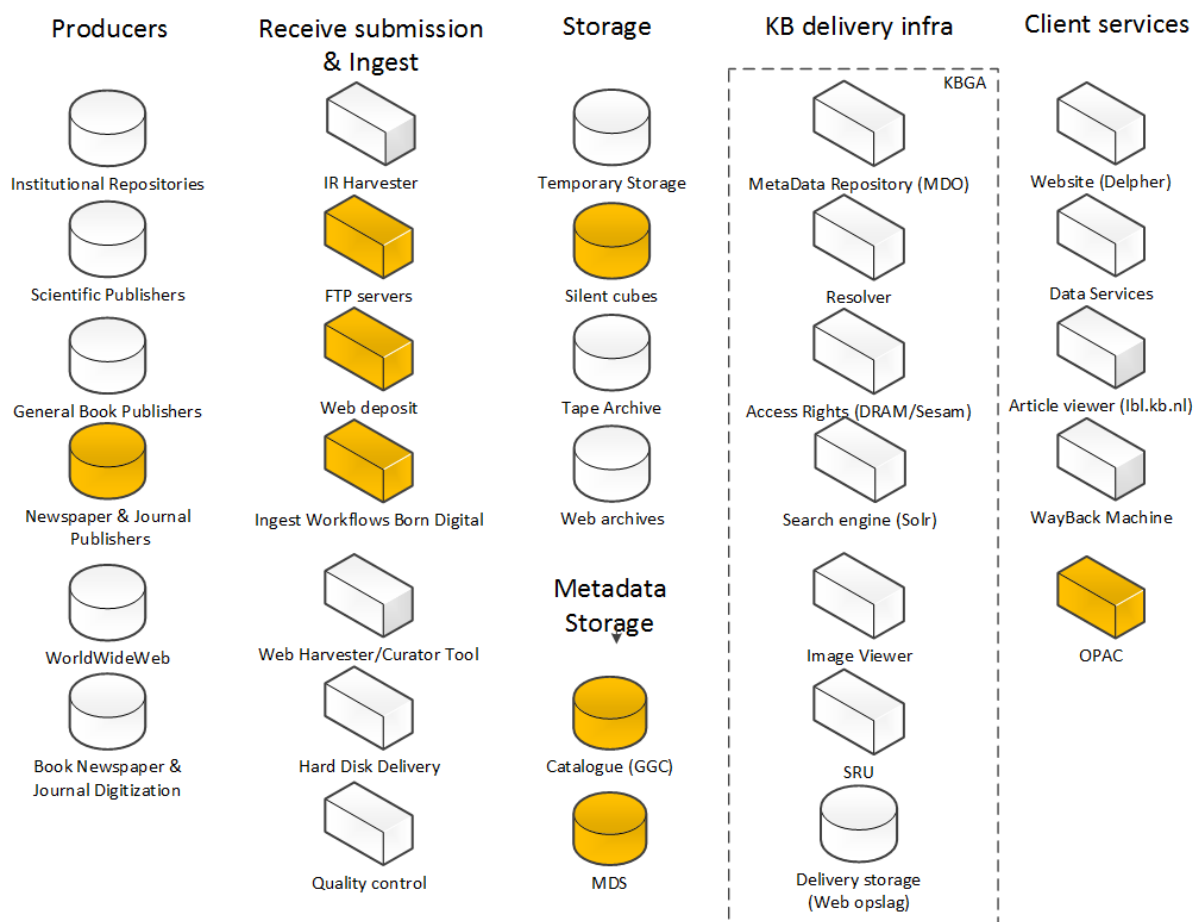
E-books can also be delivered to KB via the Monographies Digital Desk. This service is offered via the KB website. The e-book uploaded by a supplier is provided with metadata and offered on the KB FTP server for processing.

The IP to be processed comprises a PDF, metadata and annexes where applicable:

Various checks are carried out during pre-ingest, such as a 0 byte check.

E-books are made available via the OPC (library catalogue).

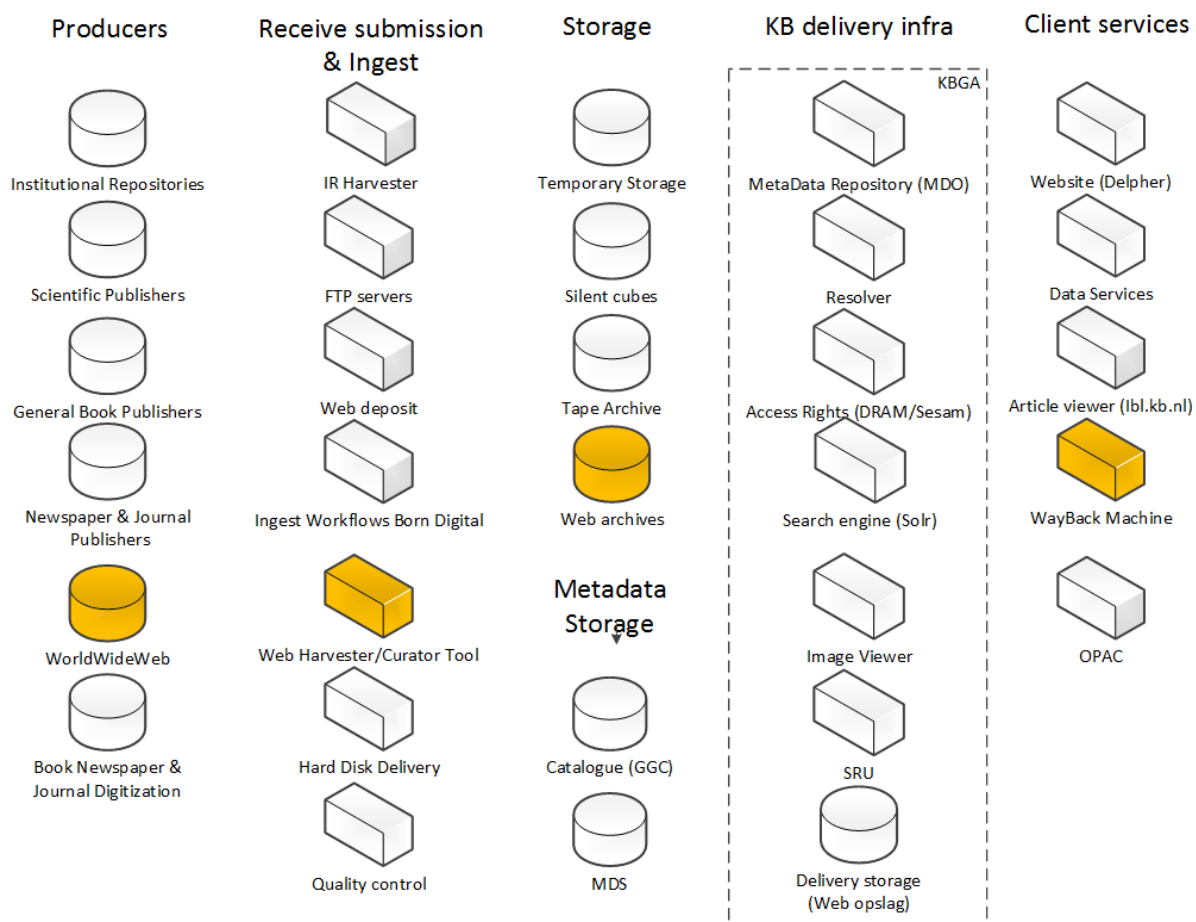
2.3.4 E-journals (periodicals)



A web service is available for Dutch publishers of online (electronic) journals (Periodicals Digital Desk; Web deposit in the drawing) which enables issues of a journal to be uploaded (in PDF format). To be able to use this service, the information about the journal needs to have been registered beforehand by ISSN Nederland, a service that is delegated to KB.

Once the registration is complete, issues can be uploaded. Metadata is created in the process and the IP is placed on the KB's FTP server for further processing. Checks are then carried out in the pre-ingest process, such as a checksum control on the files delivered and the GGC record of the journal title is added to the IP.

2.4 Web archiving



One of the duties of KB is the archiving of websites. There is currently a growing selection of 13,000 web sites that are regularly harvested (at least once a year).

The sites are harvested using the Web Curator Tool. This application is based on Heritrix and produces one or more files in .ARC or .WARC format for each harvested website. Furthermore, the application saves various log and report files for each harvest and additional descriptive and administrative metadata.

At present, the files are still stored externally (not in DM 1.5), but the aim is to include this material in DAMPS. The web archive is made available via the (open) Wayback Machine application. Access is provided in KB's reading room and part of the collection, which has IP-address protection, is made available via the Wayback Machine on the Internet in Tresoar's reading room in Leeuwarden.