

**Specifications for
Network Delivery
of**



Version 1.5

March 2006

Author: Paul Mostert

Table of Contents

1	Introduction.....	3
2	Overview.....	4
3	Structure of Net-delivered datasets.....	6
4	Datasetinfo.xml and Confirmation Email message format.....	7
4.1	Document Type Definition (DTD).....	8
4.2	Example of a datasetinfo.xml file “announcing” a dataset.....	9
4.3	Example of a Confirmation message.....	9
5	Considerations.....	10
5.1	Network transfer capacity.....	10
5.2	Security.....	10
5.3	Backing up of SDOS datasets.....	10
5.4	Unexpected difficulties.....	10
5.5	Difference between CD and Net datasets.....	11
5.6	“Downcasing” of file names.....	11
5.7	Different versions of SDOS datasets.....	11
5.8	Customer-dependent variables.....	11
6	Responsibilities.....	12
6.1	Elsevier Science responsibilities.....	12
6.2	Customer responsibilities.....	12

Version History

Date	Version	Comments
February 2001	1.0	Initial release
March 2001	1.1	Corrected some minor typos
August 2001	1.2	Procedural change of email announcement to “dataset polling”
September 2001	1.3	Corrected some minor typos
July 2002	1.4	Added download start and stop time to Subject line to Section 4.3
March 2005	1.5	Added mechanism to split .tar files over 2 Gb in size in Section 3

Acknowledgements

Thanks to Adriaan den Braber, Bryan Graupmann, Jeroen Hogendorp and Warry Spykstra, who contributed parts to this document.

1 Introduction

ScienceDirect OnSite (SDOS) datasets, which are formatted and structured in accordance with the EFFEFFECT specifications, are traditionally delivered on CD-Rom from Elsevier Science's Electronic Warehouse (EW). The delivery mechanism described in this document describes the mechanism to transfer SDOS datasets via network delivery. We refer to these datasets as "Net datasets", whereas the traditional delivery mechanism will be referred to as "CD datasets".

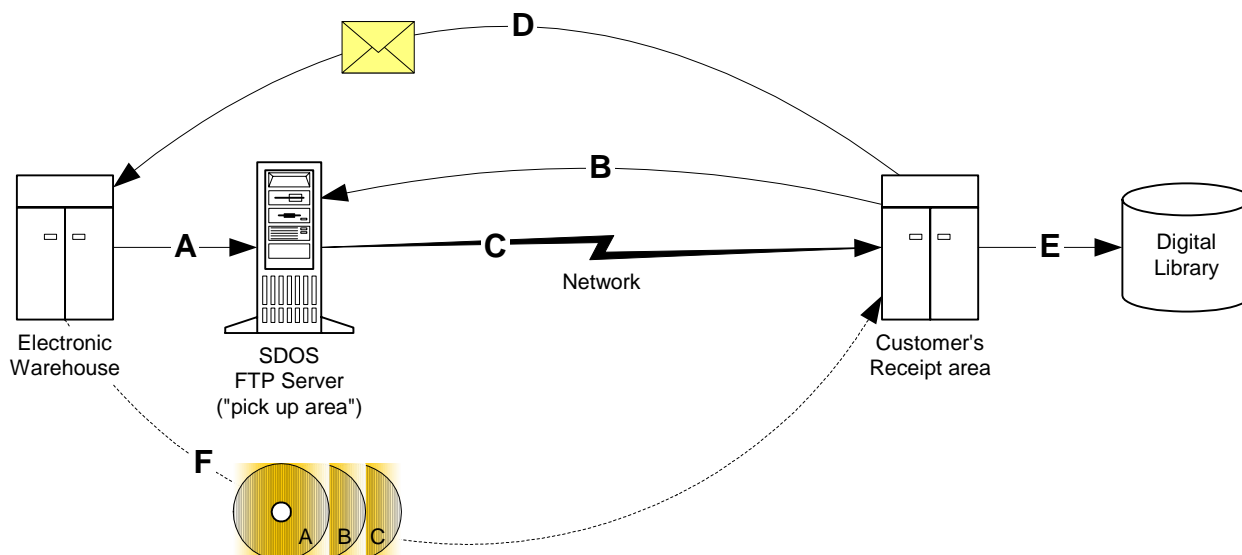
A word of caution

The systems and mechanisms for network delivery are experimental at the day this document is written. The specifications will be adapted as soon as more experience is gained, and those changes will be reflected in subsequent releases of this document.

SDOS customers that have indicated to participate in these experiments will have to develop and manage the processes at their particular sites. ScienceServer is not able to handle this for the time being, as too many variables are unknown. During the test phase these processes should be done partly manually, partly by scripts and systems developed by the different customers. Customers are invited to share successful implementations with Elsevier Science and ScienceServer LLC, to be incorporated in future releases of the ScienceServer software.

2 Overview

The procedure can be depicted in the following diagram:



The different steps in detail:

A SDOS dataset production and upload

When a SDOS dataset completes its assembly, it is TARred¹ per journal title (based on ISSN) and subsequently transferred to the SDOS FTP server in a separate dataset directory (the "pick up" area). The last file uploaded is the file *datasetinfo.xml* in the "root" of the dataset directory.

B Regular polling by SDOS customer

SDOS customers regularly (preferably via an automated FTP process script; this should at least take place daily, but preferably more often) have to poll the SDOS FTP server to see if any new SDOS datasets have been made available to them. After the upload has completed successfully, the presence of the file *datasetinfo.xml* in a dataset directory indicates to the SDOS customer that the dataset is available for download. The *datasetinfo.xml* file includes the dataset contents as an XML-formatted document. The SDOS customer is responsible for tracking which datasets have already been processed.

C Upon a successful 'poll', the SDOS customer starts download process

The availability of the *datasetinfo.xml* file should trigger the download process by the SDOS customer. The customer (preferably via an automated FTP process script) logs on to the SDOS FTP site and initiates an FTP "pull" of the tarred dataset to the designated Receipt area.

After the dataset files have been downloaded, the customer's FTP process script should compare the size and the MD5² checksums of the several received files with those from the *datasetinfo.xml* file to verify that the dataset was received intact, and that no packets were dropped during the transfer. If the sizes and MD5 checksums do not match, the customer's script should initiate a second attempt to pull the affected files from the SDOS FTP site and verify again. If the match

¹ TAR = Tape ARchive, a UNIX-based container format in which directories and files are compiled into one single file.

² MD5 = Message Digest 5, a checksum calculation format developed by RSA Inc.

fails again, the script should issue warning messages to the local system manager, who then should look into the problem and if needed contact Elsevier Science about the problematic files.

D Customer sends a Confirmation email after downloading is completed

The customer should send an XML-formatted email Confirmation message to a designated email address at the EW. The EW will use this message as a notification that the dataset may be removed from the SDOS FTP server, if this message is received within the two week “pick up window”.

Until automatic housekeeping procedures are in place, the EW only regards Confirmation email messages as indicators and will not actively approach (“chase”) SDOS customers for non-downloaded datasets. *The “pick up” area at the EW has restricted storage and system management capacity. SDOS customers have the obligation to download available datasets within two weeks, after which the pickup area is released for new datasets.*

E Processing of the dataset by the SDOS customer

After successful downloading and verification of the files, and after the Confirmation is send, the customer’s script could then initiate procedures of untarring, loading and indexing the data in the ScienceServer Digital Library system.

F Shipment of CD datasets to SDOS customers

As a fallback while performing the transition from CD to network delivery, the EW will ship CD-Rom copies of all SDOS datasets sent via FTP for a certain period until the transmission process has proven to operate stably and consistently. Those CDs will however be sent in larger shipments via lower cost delivery services than are currently required. A shipment of CDs will be made once every two weeks.

3 Structure of Net-delivered datasets

Net datasets appear as directories within the “root” of the SDOS FTP server. All files pertaining to these datasets will be available in such a directory. For instance, the directory /ohl09990 will contain all the files of dataset OHL09990.

The files within those net datasets are TARred with standard UNIX “tar” facilities. All issues of one journal title will be recursively assembled in a single .tar file that has as its name part the ISSN without the dividing dash. For instance, the file /ohl09990/00406090.tar will contain all the journal issues of the journal *Thin Solid Films* (ISSN 0040-6090) of dataset OHL09990. The .tar files are not compressed, as the majority of the contained files are already compressed according to a particular scheme, and further compression would only result in longer decompression times.

There is one exception to this rule when .tar files would be larger than two Gigabytes. This could happen, e.g., when journal issues of a newly acquired journal would have to be delivered in one “bulk” delivery. In these cases .tar files may be named with *nine* characters, ending in A, B, C, etc. instead of the normal eight characters indicating the ISSN. This indicates that a journal file has been split in several parts. For example, if there are the files 00406090A.tar, 00406090B.tar and 00406090C.tar this means that all issues for this journal did not fit in one single .tar file. When untarring these .tar files, their combined contents should be copied in the (single) folder /ohl09990/00406090 without the ninth character A, B, C.

Next to the journal .tar files in the dataset directory, the *dataset.toc* file is available as a separate file.

The last file that is uploaded is the file *datasetinfo.xml* which contains necessary information to verify the completeness and correctness of the dataset. Its presence indicates that the dataset is complete and ready for download.

An example: The directory structure of the SDOS FTP site at a given time could look like this:

```
/ohl09990
  00406090.tar
  00014575.tar
  0893133x.tar
  dataset.toc
  datasetinfo.xml
/ohl10000
  .....
/ohl10010
  .....
```

4 Datasetinfo.xml and Confirmation Email message format

The *datasetinfo.xml* file that announces the complete availability of a dataset and the confirmation email message that confirms to Elsevier Science that the dataset has been downloaded and verified correctly have the same structure as laid down in the Document Type Definition (DTD).

The following elements and attributes are required in the files/messages:

- `dataset` All-enclosing begin and end tags
- `identifier` Attribute of `dataset`, containing the name (ID) of the dataset
- `customer` Attribute of `dataset`, containing the customercode.

Note: this code could be different between the *datasetinfo.xml* file and email confirmation message in the case of fullset SDOS customers. For instance, the *datasetinfo.xml* file would contain `customer="OHL"` but CSIRO would have to return `customer="CSI"` in the confirmation message.

- `status` Attribute of `dataset`, containing the status of the dataset
- `date` Holds the date in attributes `year`, `month` and `day`
- `file` Specifies a file to download. Holds attributes `name`, `size` and `MD5`
- `name` Holds the name of the file
- `size` Holds the size of the file
- `md5` Holds the MD5 checksum for the file

4.1 Document Type Definition (DTD)

```

<!-- =====
Network Delivery Email Announcement and Confirmation XML DTD
Copyright (c) Elsevier Science 2001

=====
Version history:
- Version 1.0 27 March 2001: Initial release
=====
Typical invocation:
  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
  <!DOCTYPE dataset SYSTEM "http://support.sciencedirect.com/xml/sdosftp10.dtd">
=====
-->

<!ELEMENT      dataset      (date, file+)>
<!ATTLIST     dataset
  identifier    CDATA          #REQUIRED
  customer      CDATA          #REQUIRED
  status        (Announcement | Confirmation) #REQUIRED
  version       CDATA          #FIXED
                    "Network Dataset Announcement/Confirmation v1.0">

<!ELEMENT      date          EMPTY>
<!ATTLIST     date
  year          NMTOKEN        #REQUIRED
  month         (January | February | March |
                April | May | June |
                July | August | September |
                October | November | December )
                    #REQUIRED
  day           ( 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
                10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
                20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
                30 | 31 )      #REQUIRED >

<!ELEMENT      file          EMPTY>
<!ATTLIST     file
  name          NMTOKEN        #REQUIRED
  size          NMTOKEN        #REQUIRED
  md5           NMTOKEN        #REQUIRED>

```

The Document Type Definition is available at <http://support.sciencedirect.com/xml/sdosftp10.dtd>

4.2 Example of a datasetinfo.xml file “announcing” a dataset

Dataset OHL09990 is available for download by (fictitious) customer “Electronic University” (customer code is ELU). The *datasetinfo.xml* file could look as follows (formatted for easier reading; boldface indicates variable data):

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE dataset SYSTEM "http://support.sciencedirect.com/xml/sdosftp10.dtd">
<dataset identifier="ohl09990" customer="ELU" status="Announcement"
  version="Network Dataset Announcement/Confirmation v1.0">
  <date year="2001" month="February" day="13"/>
  <file name="00406090.tar" size="7365378" md5="64161a98b2145077eb6d9fd3421526fa"/>
  <file name="12467391.tar" size="15346720" md5="a3262f785be994e058910e67beeebef1"/>
  <file name="1146609x.tar" size="9856432" md5="df3248b579e55ada16c4b1166f3f9478"/>
  <file name="0966842x.tar" size="7340962" md5="758fad8118188beb5bacfddffe31bced"/>
  <file name="dataset.toc" size="2345267" md5="a5874a95854dc61c5ff72855e8728fbb"/>
</dataset>
```

4.3 Example of a Confirmation message

The associated Confirmation email message from customer “Electronic University” that will be sent after all files have successfully been downloaded and processed could look as follows:

```
From: sdosagent@ElecUniv.edu
Sent: Fri 14/February/2001 02:03
To: sdos_info@elsevier.com
Subject: SDOS Dataset Confirmation: OHL09990 (10:10-13:40)
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE dataset SYSTEM "http://support.sciencedirect.com/xml/sdosftp10.dtd">
<dataset identifier="ohl09990" customer="ELU" status="Confirmation"
  version="Network Dataset Announcement/Confirmation v1.0">
  <date year="2001" month="February" day="14"/>
  <file name="dataset.toc" size="2345267" md5="a5874a95854dc61c5ff72855e8728fbb"/>
  <file name="0966842x.tar" size="7340962" md5="758fad8118188beb5bacfddffe31bced"/>
  <file name="12467391.tar" size="15346720" md5="a3262f785be994e058910e67beeebef1"/>
  <file name="1146609x.tar" size="9856432" md5="df3248b579e55ada16c4b1166f3f9478"/>
  <file name="00406090.tar" size="7365378" md5="64161a98b2145077eb6d9fd3421526fa"/>
</dataset>
```

Please note the Subject line of the message in which also the start/stop times are included:

Subject: SDOS Dataset Confirmation: ohm09990 (10:10-13:40)

indicating that downloading of this dataset started at 10:10 AM and finished at 13:40 PM (local time). This makes it possible to perform statistics on transmission times and find possible network congestion bottlenecks.

In the case a transmission crosses the end-of-day, a +1 is added as in....

Subject: SDOS Dataset Confirmation: ohm09990 (23:10-02:40+1)

5 Considerations

5.1 Network transfer capacity

SDOS datasets could be very large, especially in the case of fullset customers, who subscribe to more than 1,500 journal titles. This can easily amount to datasets beyond 1 Gigabyte per working day, and in occasional peak periods about 1½ - 2 Gigabyte per day. There is evidence that the public Internet is not reliable and fast enough for these amounts of bulk data, especially to transoceanic sites, even with the large bandwidths available today.

The minimum bandwidth required would be a consistent transmission speed of a guaranteed one Gigabyte per four hours for a fullset SDOS customer (approximately 1,100 journals). This allows for overheads such as peak periods, retries/retransmissions within the same working day, resolution of errors, deadlock situations, etc. If the public Internet cannot fulfill this requirement, a dedicated network line is required in order to realize a stable, consistent, error-free and fully automated delivery mechanism. There are a number of network providers that offer so-called Quality of Service IP services or Content Delivery services with reliable throughput and response guarantees. Examples are Digital Islands, Akamai and UUnet, and more generically Internet2. The prerequisite is of course that both parties are connected to these network providers.

Given the above, network delivery of SDOS datasets will have to be treated on a case by case basis for each interested customer, as the network and other possibilities vary with geographic location.

5.2 Security

The network connection will be secured by restricting on the basis of both fixed IP addresses and username/password. Downloads will be logged. The logs will be investigated in cases of suspect or erroneous downloads. The logs will incidentally also be used for measuring transmission times on a sample basis.

Further, both parties will do their best to protect the network delivery mechanism and notify each other in the case of odd phenomena.

Elsevier Science is looking into using PKI/Digital Certificates for Extranet applications. If this appears to be relevant and/or successful for network delivery of SDOS datasets, then this will be implemented and offered to SDOS customers requesting for enhanced authentication and possible encryption and decryption techniques.

5.3 Backing up of SDOS datasets

SDOS customers receiving Net datasets do not receive physical datasets any more after the transition from CD to Net delivery has successfully completed. Where CD datasets provide backup in case of hardware/software problems at the SDOS customer, Net datasets are more “volatile”. Adequate provisions must be made at the SDOS customer to allow for backup and restoring of datasets.

In general, the EW will not delete journal issues from its archive, except because of copyright reasons (for journal titles that are sold to another publisher and the right to retain old journal issues are not granted). However, it is beyond the EW capacity to recreate datasets as a regular routine. This is restricted to “catastrophic situations” when regular backup and restore facilities at the SDOS customer emerged as insufficient. Those will be treated on a case-by-case basis.

5.4 Unexpected difficulties

There could be unexpected problems at either side of the network connection, which prevail downloads to occur. Elsevier Science will do their utmost to maintain 24 hours per day/7 days per week availability, but will not implement extensive and redundant (e.g. mirroring) capacity. If for whatever reason, Elsevier Science cannot deliver datasets via the network, or is building a too large backlog to transmit within a reasonable timeframe, datasets are shipped on CD-Rom as a fallback

until the problematic situation has been resolved (see also the next section on dataset naming conventions).

If a customer faces severe problems with network transmissions, Elsevier Science will deliver datasets on CD-Rom until the problems are resolved. The customer should notify Elsevier Science immediately of these situations to prevent any backlogs creeping up. If problems occur regularly with a certain customer, then the situation with this customer will have to be reconsidered as SDOS dataset network deliveries are offered on the basis that they can run fully automatically (including error recovery) without manual intervention.

5.5 Difference between CD and Net datasets

SDOS datasets are treated as “medium neutral” logical units. A dataset contains electronic journal material produced in a certain period (dependent of the SDOS customer), and this could extend beyond the limit of typical CD-Rom’s 650 Megabyte capacity. SDOS datasets that are too large to fit on a single CD are split in different physical datasets, each within the 650 Megabyte size limit. Typically, CD dataset names are composed of a 7-character string, appended with the single character A, B, C, etc, which denotes a subset of the original logical dataset. Each of such a physical dataset is consistent in itself. E.g., the 1.5 Gigabyte “logical” dataset OHL0999 will be split into the “physical” CD datasets OHL0999A, OHL0999B and OHL0999C.

“Net” datasets are however not split and could well extend beyond the 650 Megabyte size. The 7-character string will be the same as the CD incarnation, but the single ending character will be numeric (mostly it will be “0”). E.g., the same dataset OHL0999 will be provided as the net dataset OHL09990.

5.6 “Downcasing” of file names

For backward compatibility, SDOS datasets adhere to the ISO-9660 Mode 1 naming conventions. ISO-9660 is an international standard that defines a file system for CD-ROMs. Level one ISO-9660 is similar to an MS-DOS file system. Filenames are limited to eight single-case characters, a dot, and a three-character extension. Filenames cannot contain special characters, (no hyphens, tildes, equals, or pluses). Only single case letters, numbers, and underscores. Directory names cannot have the three-digit extension, just eight single-case characters.

All alphabetic characters in file and directory names on the CD-Rom are in UPPER case. Some software maps this to lower case. This is typically the case in many UNIX variants, in which the CD-Rom driver translates the uppercase directory and file names to lowercase by default. However, Net-delivered dataset file and directory names in *.tar* containers will not automatically be “downcased” in the untarring process. However, the *.tar* files themselves and the *dataset.toc* and *datasetinfo.xml* files will be downcased.

This note is only relevant for non-ScienceServer customers, as the ScienceServer system performs automatic downcasing of directory and file names in the loading process.

5.7 Different versions of SDOS datasets

The approach described in this specification is “neutral” with regard to SDOS versions. All versions of ScienceDirect OnSite (EES v1.0-v1.2, SDOS v2.0 and v2.1, the future SDOS v3.0, and CITADEL v1.0) can all be handled with the delivery mechanism described in this document.

5.8 Customer-dependent variables

There may be slight variations in the set up of network delivery for a particular customer. For instance, initially the system is envisaged to run from Amsterdam. However, Elsevier Science may decide to install different FTP servers at different locations in the world to optimize network traffic to certain customers in different geographic locations. Therefore the IP address of the FTP server is not fixed. The email addresses for Confirmation messages may also vary per customer. Those will be communicated separately.

6 Responsibilities

6.1 Elsevier Science responsibilities

- Make available SDOS datasets in a timely manner.
- Clean up FTP site after Confirmation messages were received that customers have successfully processed datasets, or after two weeks, whichever comes first.
- Elsevier Science takes necessary precautions to ensure security.

6.2 Customer responsibilities

- “Poll” regularly (at least once per day, preferably more often) for new SDOS datasets that have become available for download. It is strongly advised that the customer develops an automated script (e.g. a “cron” jobs) to perform this task.
- Customer starts downloading a dataset within 24 hours after availability, at a 24 hour/7 days a week basis. It is strongly advised that the customer develops an automated script to perform this task.
- Customer verifies that a daily download can be performed within 24 hours, with ample spare capacity for error recovery and peak periods. To this end, sufficient bandwidth from his or her own location to the SDOS FTP server must be arranged. Elsevier Science will provide assistance in this on a case-by-case basis.
- Customer emails XML-formatted Confirmation messages to the EW immediately after the download and verification have finished successfully, and before loading and indexing processes are started.
- Customer verifies that the sizes and MD5 data in the Announcement message match with the actually downloaded dataset files prior to further processing, and performs at least one (automatic) re-download. If size and MD5 still fails to match, the customer notifies the EW directly, or the associated Elsevier Regional Sales Office via traditional channels.
- Customer takes precautions to secure downloads and notifies Elsevier Science if unexpected events occur that could indicate security breaches.
- Customer implements adequate backup and restore procedures in the case of system failure or other emergencies. Note that CD’s are not provided any more after the SDOS has switched to network delivery.