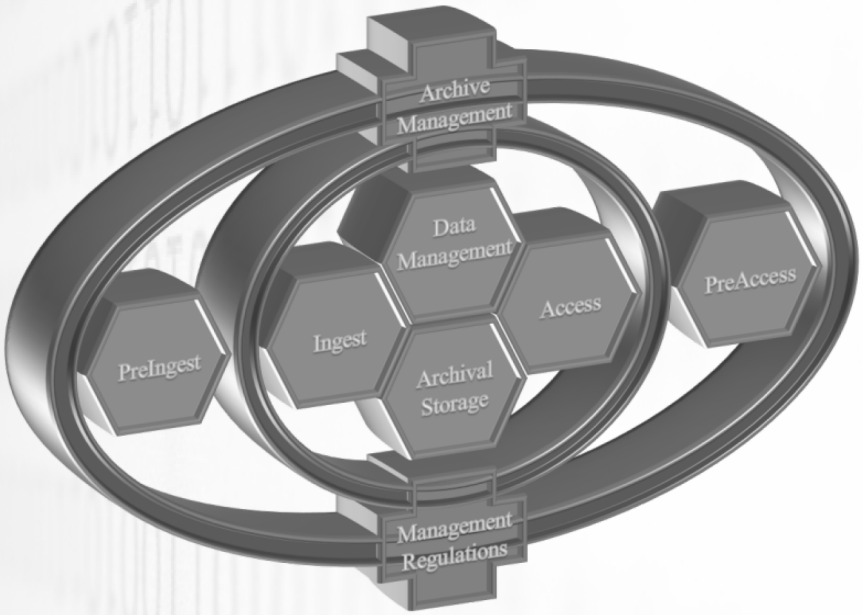
ESSArch

ESSArch Tools and ESSArch Preservation Platform



This document provides a basis for understanding ESSArch and how it can be used for digital long-term preservation.

It describes the use of ESSArch Tools within organizations producing information to be archived as well as within organizations preserving digital information.

Furthermore, the ESSArch Preservation Platform is explained and how it can be used to secure integrity and long-term preservation of archived information packages. How this is performed when information is to be stored, preserved and made available and accessible to authorized consumers.

ESSArch is based on OAIS and licensed as Open Source.

TABLE OF CONTENTS

1. ABOUT THIS DOCUMENTATION

1.1 Purpose

This document describes ESSArch which consist of ESSArch Tools (ET) and ESSArch Preservation Platform (EPP).

1.2 Terminology

Terminology used in this document is from the conceptual model OAIS[1].

1.3 Reading instructions

The document provides a basis for understanding ESSArch and how it can be used. There are different areas of focus:

- Generally about ESSArch – overview of ET and EPP
- ESSArch in a e-archive production environment – functionality and relationship between ET and EPP
- Delivery to and from an e-archive – different zones explained
- Scenarios in a e-archive production work flow – a scenario of delivery to and from a digital e-archive
- Architecture of ESSArch – overview of system architecture, code and directories

1.4 Versions

Date	Rev	Signature	Information
2013-04-30	1.0	ES Solutions	Original
2014-02-24	1.1	ES Solutions	Updated in conjunction with the project E-ARK

[1] Open Archival Information System, ISO 14721:2003

2. GENERALLY ABOUT ESSARCH

ESSArch consist of Tools (ET) and a Preservation Platform (EPP). Together they support the whole process when information are structured and packaged as SIP's, delivered to a preservation platform, stored as AIP's and made accessible as DIP's. Together they bring cost effective functionality for creating and managing archived information. ESSArch is a multi-platform licensed as Open Source. ET is distributed as Linux and win32/64 packages and EPP is distributed as a Linux package.

2.1 Short description of ESSArch Tools (ET)

ET is briefly a SIP package tool with logging (eq. notes) capabilities. It provides mechanisms for preparing, creating, delivering and receiving SIP's and along the way creating manually notes about the steps taken. It uses METS to describe SIP content and SIP packages (TAR-files) as well as PREMIS for content preservation. Notes are stored as PREMIS events. The SIP metadata (content/package/notes) are described as xml-files and based on the specifications for SIP packages used in Sweden and Norway. Both xml structure and the physical content represented in the xml structure are validated during creating process as well as when receiving SIP's. ET can be installed as windows 32/64 binaries and on Linux. A basic installation of ET is profiled as a producer (OAIS terminology) but can easily be switched to a receiver of SIP's as an archival institution (PreIngest/Ingest, OAIS). ET can also be profiled as being used within highly secured environments with only logging capabilities. All these three profiles can easily be switched within the application. ET is a stand-alone application and can be used as a complement to EPP.

ET can be used by those who produce information to be archived as well as by organizations which receive and preserve information. ET is configurable and adaptable to the processes and procedures that exist within a producer organization as well as within a preservation organization (blue/orange/red zone). All these zones are explained further on in the document.

2.2 Short description of ESSArch Preservation Platform (EPP)

EPP is briefly a preservation platform. EPP can manage IP's in a controlled environment (control area) and save them to archival storage environment. A SIP can be checked-in to the control area, as an original IP and then be checked-out to a work area for various operations and afterwards be checked-in again to the control area. EPP manage different generations of IP's in the control area. The archival information collection, AIC (OAIS terminology), is used for generation handling of IP's. Within the control area a differential check can be made between generations of processed IP's and its associated original IP (SIP). This can be done before IP's can be approved for saving to archival storage. An AIP is described with METS for both content and package (TAR-file) and by PREMIS for preservation metadata. Exchange of information, eq. different request types, with an AIS through an one-way gateway can for example request EAD/EAC descriptions, place them on the work area for further processing and then be checked-in as an AIU (OAIS terminology) and saved as a new AIP with related metadata added to AIC. This means we classify/type different IP in the archival storage environment. Every registered event (note/log) for each processing IP, from the producer (prepare/create/deliver/receive etc) until created as an AIP are saved as a PREMIS preservation file to archival storage. An access request could be handled in different ways. For example, a DIP request can be handled in a way that EPP executes the request and provides an archived AIP as it is (packed) or unpacked to the work area, for further processing. EPP also provide functionality for reports, statistics and management by different configuration and policy options.

EPP can be used by the organization which produce information to be archived as well as by organizations which receive and preserve information. EPP is configurable and adaptable to existing processes and procedures in organizations.

3. ESSARCH IN A E-ARCHIVE PRODUCTION ENVIRONMENT

3.1 Information and Records Management

A digital production workflow where the primary task is to preserve the information so that it is useful at any given time, both for short and long term, requires a good long-term information management. That leads to as early as possible prepare and facilitate the preservation of information. Support, guidelines, regulations and laws that regulate and counsel in these contexts are available to accomplish this.

According to ISO 15489 [2]a record can be created and managed, whatever its technical format. Across all formats and systems it must be securely retained and managed for defined periods of time in accordance with an organizational policy for a variety of reasons – for example:

- to meet legal and regulatory obligations
- to provide for efficient reuse
- to satisfy historical and business interest

As shown in the figure below, records are usually created or received by line-of-business systems, such as document management systems, email and other communication systems, and specialized applications such as finance and ERP systems. These records have traditionally been captured into and managed by a separate records management system although suppliers are also increasingly supporting record management within their business systems.

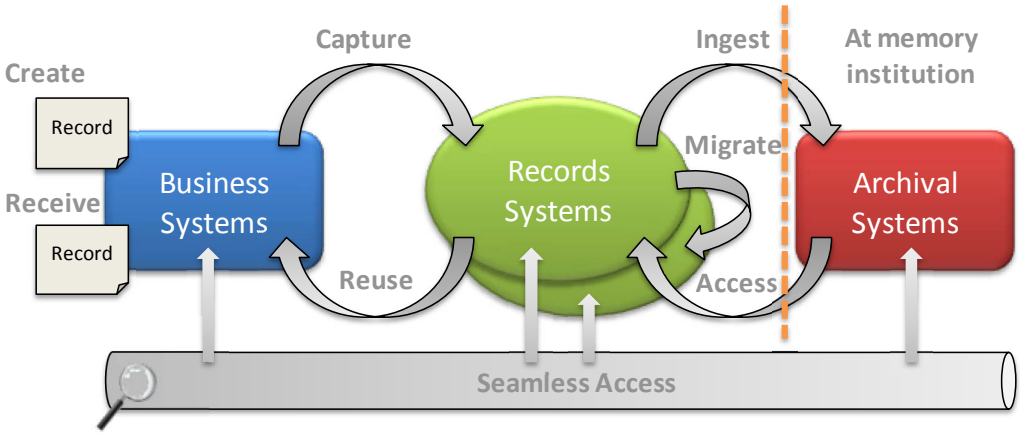


Figure 1: Seamless information- and workflow for Records Management Systems

Records may have to be transferred either individually or at specified intervals from business systems to a separate records management system. Records that have to be retained for long periods (or indefinitely) may be transferred to specialized archival systems. Records may also be migrated from one records system to another as organizations restructure, split and merge, and as older technologies are upgraded or replaced.

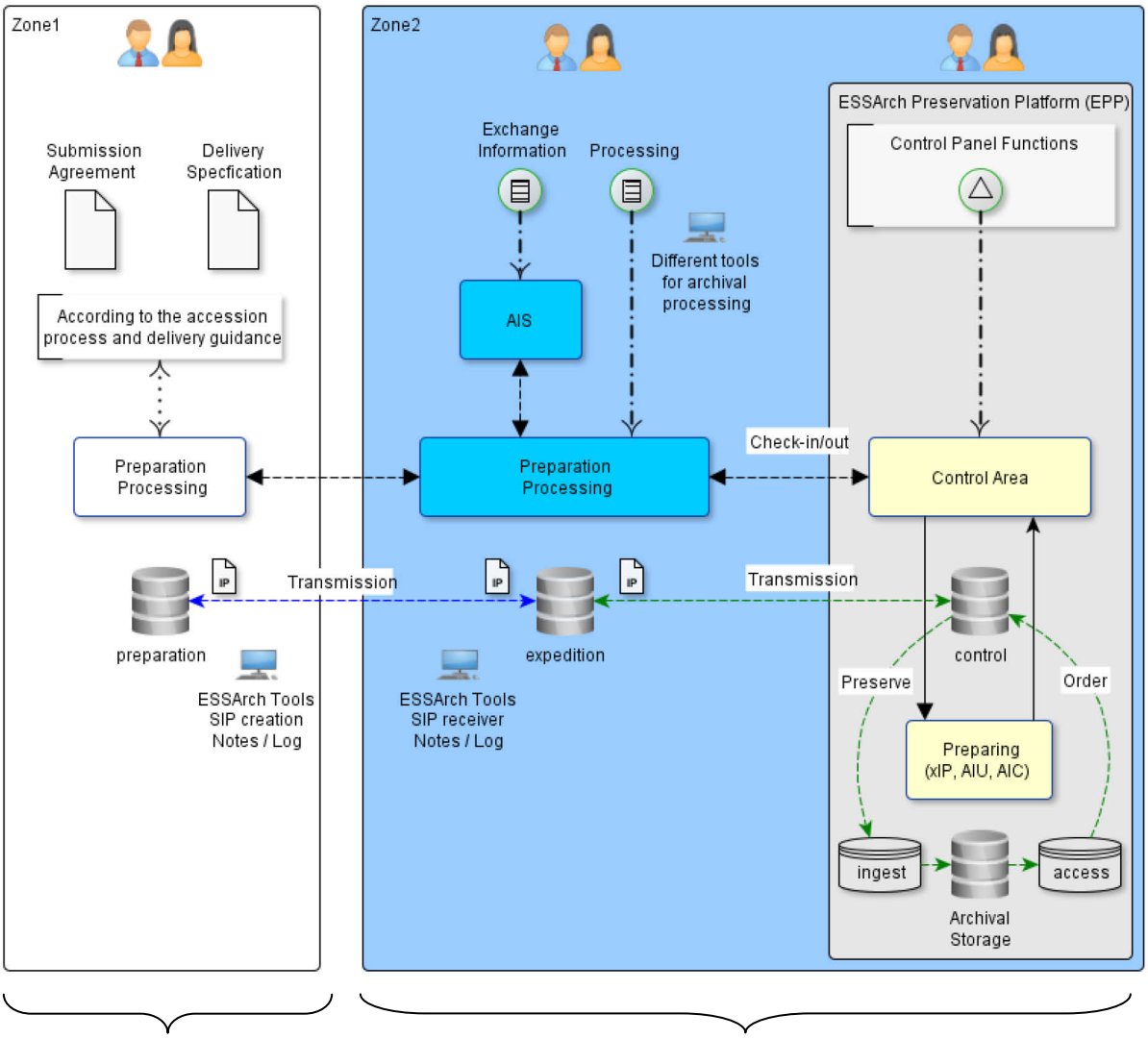
The movement of records between systems may occur many times during their lifespan and requires robust interoperability between those systems. A typical records management system contains records arranged into aggregations, metadata relating to records and other entities, a business classification scheme, a set of retention and disposal schedules, user access controls and definitions, a search engine and so on. An archival system preserves archived records and information about them, in order to be retrieved over time.

The credibility of information received, preserved and made available should not be questioned. This means an increased need for mechanisms to over time ensure information integrity and to control access to the influence of the preserved information. The digital production workflow for an archival system could be described by three different, yet similar, profiles.

[2] ISO 15489 – Information and documentation – Records management

3.2 Profile 1 – ESSArch from a producer and consumer perspective

ESSArch (ET and EPP) is provided and considered to be adaptable in different organizations with different requirements. But the principals of information workflow in an archival system should be transparent and vendor independent as well as usable in any kind of technical environment. From the perspective of a producer and a consumer of archival information there should be no differences in the way they ingest and retrieve the information. An organization can also be both producer and consumer.



A producer (zone1) can use ET to:

- create information packages (SIP's) for delivery to e-archives
- log events related to the information packages (SIP's)

A preservation organization (zone2) can use ET and EPP to:

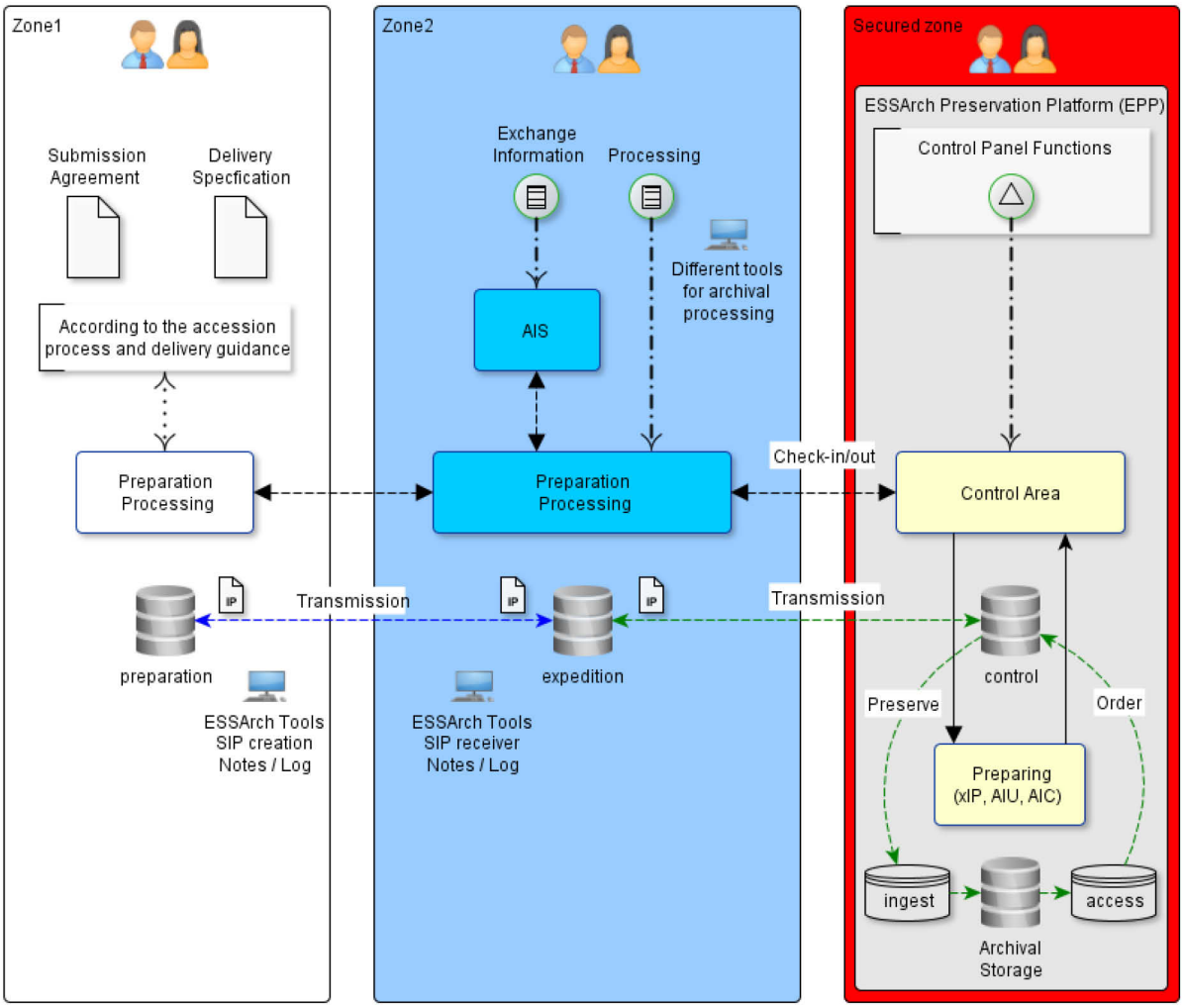
- receive SIP's and prepare them for further processing in EPP
- create SIP's of previously non-packaged received deliveries
- manage IP's in a controlled environment, providing check-in and check-out capabilities to the control area
- create AIP's for archival storage
- provide access to AIP's and make them accessible as DIP's
- log events related to all processing and IP's

A producer prepares, creates and delivers information packages (SIP's) according to established descriptions of the accession process and delivery guidance.

A preservation organization, for example an archival institution, receives information packages (SIP's) validates and process them and stores them into appropriate storage architectures. Furthermore, the organization makes archival information packages (AIP's) available and accessible to authorized consumers.

3.3 Profile 2 – ESSArch in a secured environment

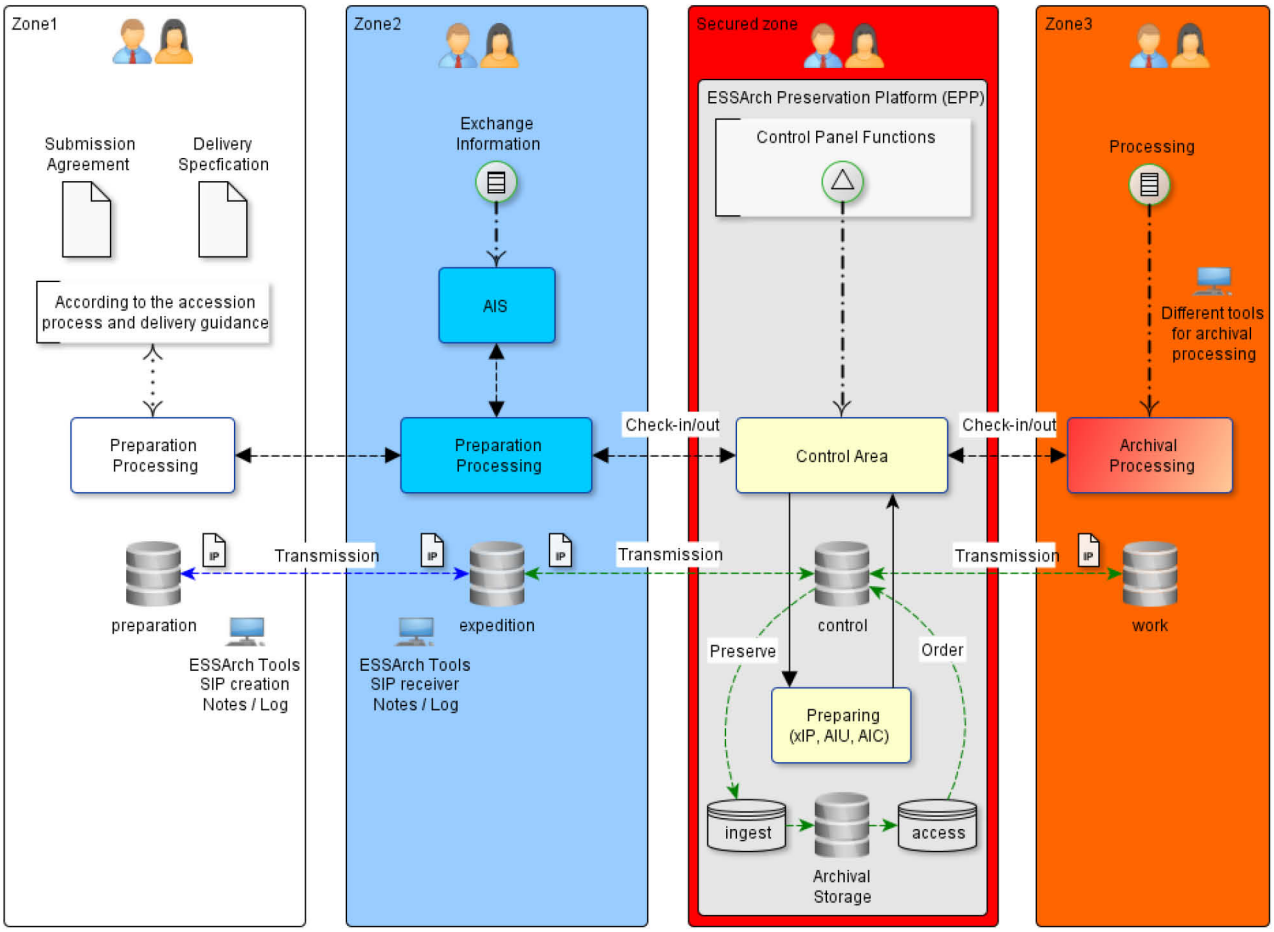
The difference from profile 1 is mainly the separation of EPP from zone2 into an isolated secured zone. The main purpose of this is to separate and isolate the preservation platform, both from an infrastructure and architectural perspective, but also to separate the management of restrictions of access, in order to secure the environment where the archived information is stored.



Once you are authorized to access the secured zone and EPP you can by EPP access and check in or out result of necessary interaction between zone2, which could be public, and the secured zone, which is isolated and secured.

3.4 Profile 3 – ESSArch in a secured environment with post archival processing

The difference between profile 3 and profile 2 is mainly the separation of archival processing into a separated zone3. The benefit of doing so is that information checked in to EPP can, at any time, be checked out to a work area, be processed and then checked in again to EPP. All information packages are stored in the control area, an environment controlled only by EPP, until the information packages are approved and prepared for archival storage, the final destination for the archived information packages.

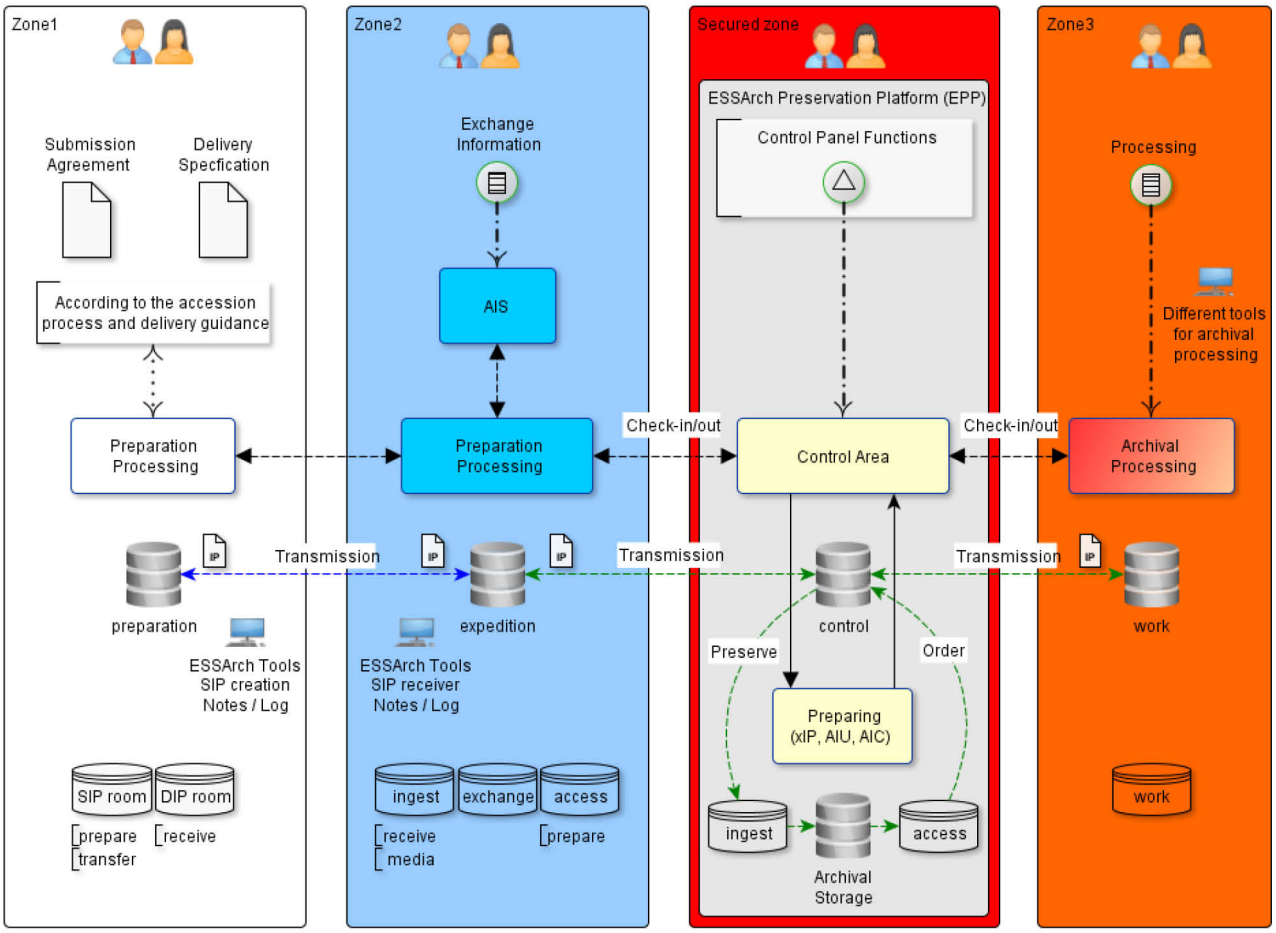


Archival processing could include the creation of DIP's as well as preservation care taking of any stored AIP. Generations of IP, received as SIP's, retrieved as AIP's or DIP's, or complemented AIUs are all structured and maintained by AIC's. All these acronyms can be found in OAIS. The workflow is however something that can be agreed upon from time to time and by different stakeholders. It is normally declared in a submission agreement.

4. DELIVERY TO AND FROM AN E-ARCHIVE

A delivery to and from EPP will be under surveillance and control, continuously. It is, as mentioned before, important to secure the credibility of information received, preserved and made available so it not will be questioned. This means an increased need for mechanisms to over time ensure information integrity and to control access to the influence of the preserved information. The basic workflow contains four zones.

- A white zone where the information to be preserved is produced and packaged, and where the producer (creator) is responsible for the information.
- A blue zone, a public expedition at the preservation organization where the access and influence of information is regulated. The preservation organization is hereby responsible for received information.
- An orange zone, not public and where the access and ability to influence the information is highly restricted and controlled.
- A red zone, not public and where access to the zone is strictly limited. Ability to influence the information is regulated to only a few functions.



Information to be preserved will be packaged by the producer (creator) and delivered to the expedition in the blue zone where a number of controls including virus control will be processed. The archival information system (AIS, not included in EPP) will be updated with the accurate details. The delivery, SIP, will then by authorized personal be checked into the control area in the red zone. The SIP can then be checked-out into a work area for testing and further processing and then be checked-in again to the control area. When adequate preparations have been made, decisions about preserving the IP's can be taken and then IP's will be stored in accordance with established structures of the archival storage environment. Preservation has thus begun. Accessing an archived information package, AIP, is usually initiated by an inquiry or request. The request could be processed in the orange zone and requested data is produced in the red zone. The requested data, DIP, is if necessary, dispatched to the blue zone, otherwise to the work area for further processing before dispatching.

4.1 White zone

The producer who wants to preserve information prepares and creates information packages according to the accession process and delivery guidance. This is done by using ET. The created information package, SIP, is then delivered to the expedition in the blue zone within the preservation organization. Package description for the SIP is sent by e-mail to preservation organization.

4.2 Blue zone

The information package, SIP, is received at the expedition and registered as received in different systems as well as in the Archival Information System (AIS). A receipt is sent in return to the producer. The SIP then undergoes different quality checks, for example virus control (not performed by ESSArch) and consistency checks and validation. Meantime these actions are performed they are also logged manually as events in a log circular (log file). It is now important and further on to keep the SIP intact and not to influence the SIP. The integrity of the SIP shall not be questioned. All the processing results complements the existing SIP and everything is finally saved to a so-called data-lock (exchange area). The data-lock is used for information exchange between the blue and orange/red zone. When everything is done at the expedition the SIP is checked in to EPP and the control area in the red zone.

4.3 Orange zone

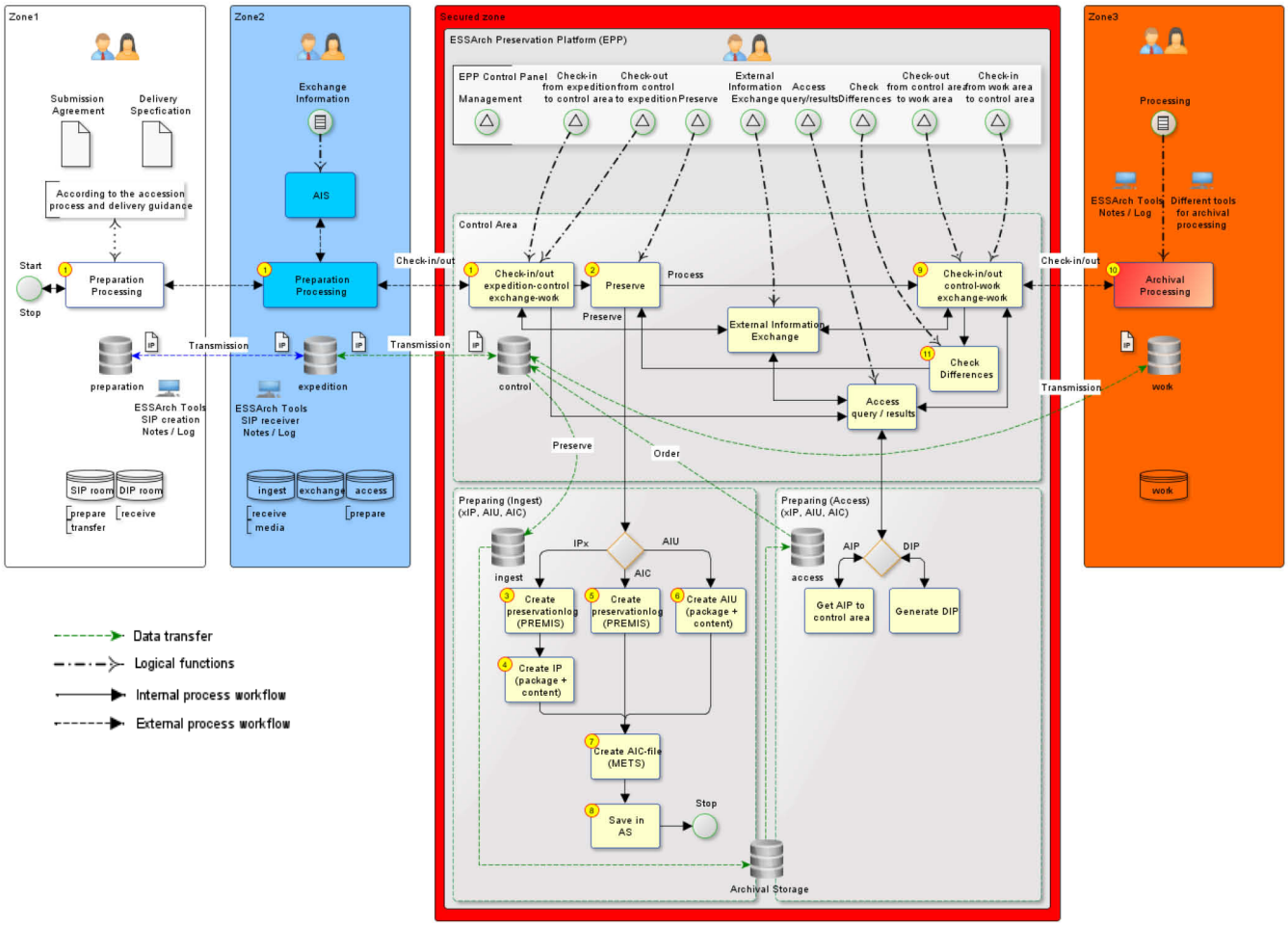
The work area in the orange zone is used for test and archival processing purposes. An IP can be unpacked, metadata can be added, deleted or changed, converting of file formats and request for resending of IP can also be performed, to mention some of the processing that can be done in the work area. All these processing can be manually logged in the intended log circular. When everything is processed a check-in to the control area can be done.

4.4 Red zone

When a request for check-in to control area is executed a compilation of SIP and different results from the data-lock (exchange area) is made and the result (IP_0) is placed in the intended area within the control area. Every processing step taken is being validated according to chosen configuration. When the SIP has been checked-in it can either be saved to archival storage or checked-out to work area for further processing. If checked-out to work area it can be restructured etc. and afterwards be checked-in again to the control area. When the IP has been checked-in again it should be considered to be a processed IP and therefore be named IP_1. We now have the original IP (SIP=IP_0) and the processed IP (IP_1). It could be interesting to know the differences between them and therefore a diff check is performed. The purpose of this diff check is to verify all processing being made in the work area and earlier in the delivery process. It is also important to verify this for later approval of IP's in order to save them to archival storage. Whenever an IP is approved it can be stored into archival storage. When this is done, preservation has started.

5. SCENARIOS IN A E-ARCHIVE PRODUCTION WORK FLOW

In this document we define a scenario as a normal sequence of events. Deviations are not explained. All events are being logged by ET in intended log files or in the database of EPP.



There exist many different kind of sequences of delivering information packages (SIP's) to a digital e-archive (TRAC[3]). It is often based on the requirements of automation and different responsibilities in the organizations. It is also affected by different preservation strategies. All this should be agreed upon in a Submission Agreement. The scenarios described are sequences where quality checks and validation are performed in a controlled environment and where certain requirement has to be fulfilled. However, regardless sequences, all received SIP's will finally be stored as AIP's in the digital e-archive. Access to archived information can also be provided in many ways, for different purposes. The scenario is described as a profile 3 scenario.

Of course there will be deviations within the sequences. These are taken care of as far as possible. Different deviations require different actions. If a deviation is classified as critical or as an error certain actions need to be considered and they could be manually or automatically executed. It is not possible to approve IP's for archival if not all deviations have been taken care of. The degree of severity and actions is determined case by case. Logging is done for both system and process workflow. Events are registered in physical files and in the EPP database.

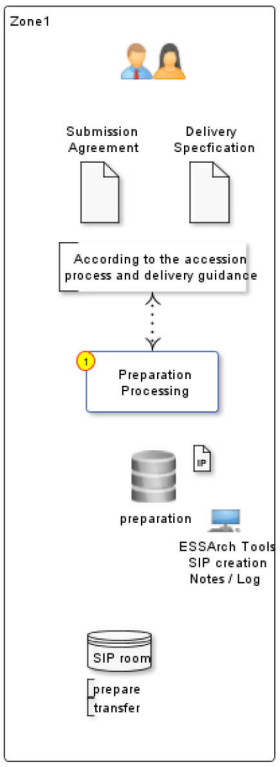
User management includes users, roles and permissions and it manages what can be done by whom in the workflow. You always have to be logged in as an authorized user to perform tasks.

[3] TRAC - Trusted Repositories Audit & Certification, originally sponsored by RLG and the US National Archives and Records Administration, the work grew to incorporate and leverage work from several organizations, laying the groundwork for international collaboration on digital repository audit and certification between the DCC, RLG (now OCLC-RLG Programs), NARA , nestor, and the US Center for Research Libraries

5.1 Scenario – Delivery to the e-archive

The scenario describes a delivery of information packages (SIP's) from a producer to a preservation organization for digital long-term preservation in an e-archive.

5.1.1 White zone – Producers, deliveries and information packages



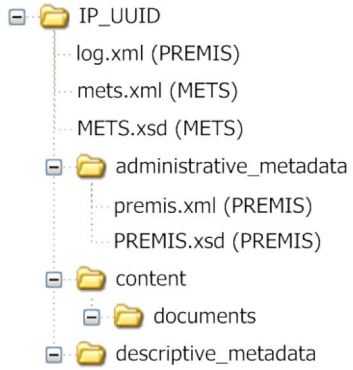
Information to be preserved needs to be prepared for this according to agreements with the preservation organization. In order to facilitate receiving, preservation and availability of information supplied.

Preparations for delivery includes creation of an information package, SIP, with a fixed content description *mets.xml* and a package description *info.xml*.

Information packages can be created by ET as TAR-files with associated checksums or as an open directory structure. A TAR-file checksum is described in the package description *info.xml*. Afterwards the package is saved and delivered either on appropriate medium (carrier) by courier/mail or by any other transmission technique like ftp etc. The preservation organization then receives the SIP for quality checks.

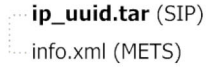
The package description *info.xml* can be sent by e-mail to the preservation organization.

An information package (SIP) can be created in a directory structure like the one described below. If the deliver consist of only one SIP, as an open directory structure, the package description will be represented by the content description *mets.xml*. If the deliver consist of one packed SIP or several packed SIP's a package description *info.xml* will be created. Events related to a deliver, SIP, will be manually registered in ET and saved in enclosed log file *log.xml*. ET is also used to prepare, create and deliver SIP's. Content in SIP's are also described by *premis.xml*, as preservation metadata. Associated schemas (xsd-files) for used metadata description files will be stored in the SIP.



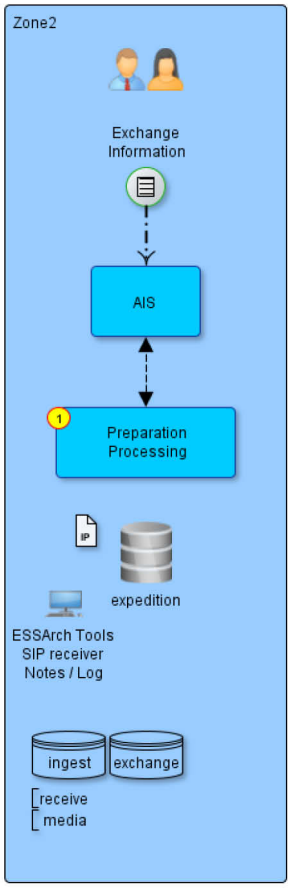
IP_UUID – unique directory for information package (SIP)
log.xml – log file
mets.xml – content description for SIP
METS.xsd – schema for METS
administrative metadata – producers metadata for "content"
premis.xml – preservation metadata description file
PREMIS.xsd – schema for PREMIS
content – delivered content on medium/carrier/package
documents – complementary documentation for "content"
descriptive metadata – producers metadata for archival description

If an information package (SIP) will be delivered in a container format it will be packaged as a TAR-file with a unique identifier and described in the package description *info.xml*.



ip_uuid.tar – packed information package (SIP) with unique identifier
info.xml – package description for the delivery (SIP)

5.1.2 Blue zone – Receivers preparation for processing



The delivery is stored in the expedition area (ingest-receive/media) when it first arrives in the blue zone. The SIP's are often stored on medium/carrier such as CD/DVD or on a USB memory stick and the content can be copied to the ingest media area.

The package description is normally received by e-mail and it is initially used for registering enclosed information into journal systems and archival information systems. After this is done, a receive receipt is sent to the producer.

An information package is identified by its unique identifier (IP_UUID) found in both *info.xml* and *mets.xml*.

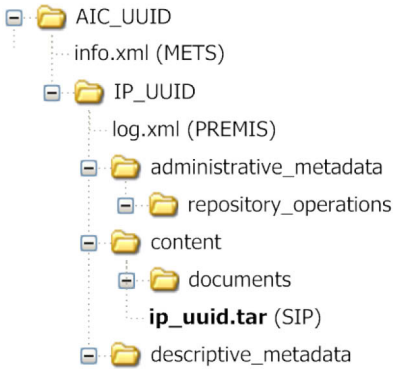
The content description *mets.xml* represents all the physical files in the SIP as well as associated checksums.

The SIP, shall under no circumstances be influenced after it has been received and shall be intact in further processing.

When the SIP is about to be interpreted it also will, at the same time, be declared as received by the preservation organization. This is true if the SIP passes validation of its structure and content. If not, a notification must be sent to the producer, requesting a retransmit of the SIP. If the SIP passes validation it can be delivered directly into the AIP creation process or delivered to the secured control area.

If the SIP will be undertaken delivery control tests it will also be transferred to a new directory structure on the data-lock (exchange) in the expedition area. The new directory structure is similar to the one used for SIP creation except that it has a parent directory with its own unique identifier (AIC_UUID). This parent directory is an AIC (OAIS terminology) directory for IP's and it will be used in the next process step. The SIP will be copied into the parent directory (AIC_UUID) on the exchange area. A new log file will also be created by ET in order to register events that occur within the upcoming processing in the preservation organization. This log file *log.xml* will be stored at the data-lock (exchange) under associated IP directory (IP_UUID). The package description for the SIP, *info.xml*, will be found directly under the AIC directory root.

Directory structure data-lock (exchange)

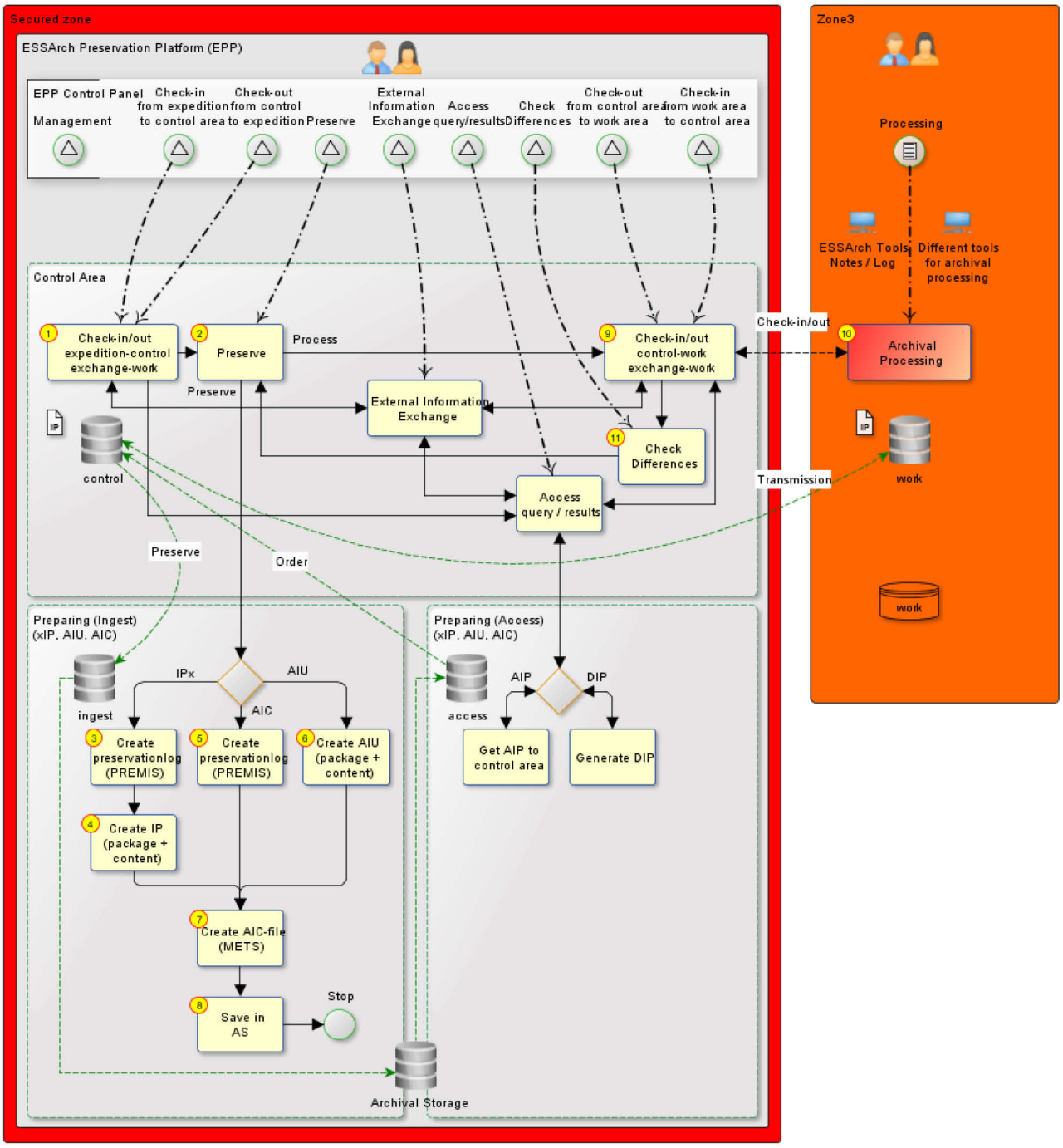


AIC_UUID – unique directory for several IP's related to each other
info.xml – package description for SIP
IP_UUID – unique directory for information package (IP)
log.xml – new log file created with events related to receiving SIP's
administrative metadata – producers metadata for "content"
repository_operations – preservation complementary information
content – the delivery, e.q SIP, packed TAR-file
documents – complementary documentation for "content"
ip_uuid.tar – information package (SIP) with unique identifier
descriptive metadata – producers metadata for archival descriptions

Virus control is then done on the delivery and a virus control report is created and added to the associated IP directory (IP_UUID) at the data-lock (exchange area). Events are manually registered by ET.

If everything has been checked and is in order the next step is to check-in the delivery for further processing.

5.1.3 Orange and red zone – Receivers processing and preservation of deliveries

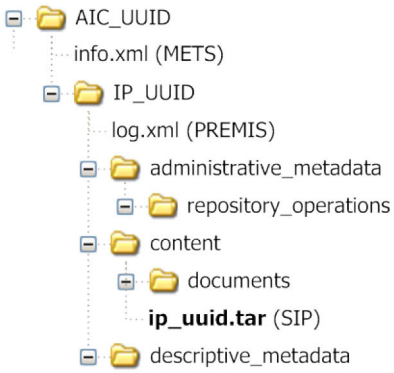


1. Check-in/out to control area

A check-in to control area in the red zone is activated by an authorized user in EPP. The deliveries (SIP's) found on the data-lock (exchange area) in the blue zone are presented to the user in the EPP GUI. When a SIP is selected it is compiled with the associated SIP from the data-lock (ingest-receive/media) and the result is placed in the control area, addressable by the SIP associated AIC used on the data-lock (exchange). To make this compilation possible, the SIP in the ingest-receive/media area must be identified with corresponding SIP on the exchange area. To place SIP's on the exchange area is something ET does when it is receiving SIP's in the blue zone (see previous chapter).

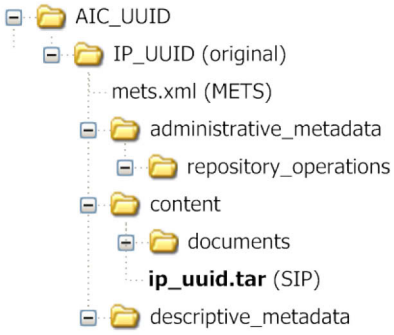
In detail, the package description *info.xml* contains a unique identifier for the SIP, the same one used in the filename of the SIP as for the directory structure on the exchange area. This directory structure belongs to a parent directory which has a unique identifier, AIC_UUID. The parent directory and all its subdirectories etc are copied into the control area. After that the content on the ingest-receive/media area will be copied into the subdirectory 'content' in the control area, if it does not already exist.

Directory structure on exchange area



Structure ingest-receive/media area

Directory structure in the control area



Original IP is the delivered SIP and the unique identifier is inherited to represent the original IP within the control area. During check-in a new content description *mets.xml* is created based on the fact that we now have a new directory structure with content that consist of delivered SIP.

Delivered SIP

When check-in is executed the package description *info.xml* is interpreted and some information is used for the creation of a new content description *mets.xml.* Some information also defines different types of check-in requests, for example, if the SIP shall be manually approved for preservation and/or further processing or if a SIP from ingest-receive/media automatically shall be preserved according to defined policy and storage method. To summarize, different request types are interpreted from *info.xml* and executed. Afterwards the package description *info.xml* has completed its mission and is no longer needed.

Registered events associated and enclosed with the SIP and the events registered by the receiver at the preservation organization and placed on the associated SIP on the exchange area is now joined and interpreted and the result is saved into the EPP database. Therefore, we do not need any physical log files because we use and store events in database when processing in the control area. Basically, the need for physical log files only exist when we do not have access to the EPP database which is when we are outside the control area.

When a check-in to control area is completed a new content description *mets.xml*, is created representing the new IP_0 entitled 'original IP' and it is also prepared for relationship with other IP's. All those relationships are handled by an AIC description stored in the EPP database.

We now have received the SIP, saved it in a controlled environment, heavily reduced the risks for influence of the SIP and entitled it IP_0 and we are now able to hand it over for further processing

2. Preserve

In the control area several IP's will coexist. If an IP shall be saved to archival storage as an AIP, an approval is required. An IP can also be checked-out to work area for further processing, though it is not to be archived and therefore not require an approval. If the original IP_0 is to be processed in the work area and checked-in again to the control area it will be entitled and considered as IP_1, a 'processed IP'. This processed IP can differ in many ways from the original IP. It is therefore important to detect differences so that actions can be taken if necessary. An approval requires that all differences and deviations have been checked and been taken care of. If not then it might not be approved for preservation in archival storage.

3. Preparation for ingestion to archival storage (step 3-8, SIP-AIP conversion)

When the IP is approved for preservation/archiving it enters the preparation stage where the IP will be described in a similar way as it was when it was addressed as an IP_X but with the difference that associated information stored in AIC also will be archived. An IP will be archived as an AIP and the AIC information will be exported from EPP database and stored as a METS-file. Together with the AIP and AIC there will be an exported PREMIS-file of all the associated events stored in the EPP database saved as an AIU. The minimum of what is being saved to archival storage are three objects for every IP saved.

Preparation also concerns preservation planning such as how many sets of an AIP shall be stored on how many medium/carriers or in different storage environments. EPP uses different policies for this purpose, configurable when the SIP is checked-in for the first time or as early as when the SIP is created by the producer. EPP uses a special I/O engine to control all the sets of archived objects, to different storage environment, and in harmony with preservation metadata in the EPP database.

When the IP has been approved for archiving and later on archived and safely stored in the archival storage it must be available and accessible to consumers. This is explained as another scenario.

9. Check-in/out to work area

Check-out to work area

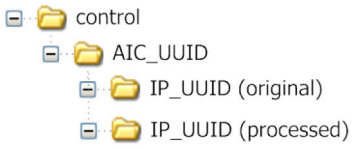
A check-out to work area is activated by an authorized logged in user in EPP. A list of possible IP's to check-out can be found in the control area. A selection of IP's has to be done if they are to be checked-out. When selected IP is being checked-out a new unique identifier is created for a processed IP. This unique identifier is added to the associated AIC in the EPP database. The main purpose with this operation is to later on be able to separate the original IP from a processed IP being checked-in again to the control area.

When selected IP's are being checked-out they are copied with their parent directory, AIC_UUID, to the logged in user home directory on the work area. If the IP's are compressed they need to be uncompressed to intended directory. Processing can begin afterwards.

Check-out events are registered in the EPP database and a new log file *log.xml* is created in the root of processed IP in the intended directory on the work area. All previous events for associated IP's described by AIC are exported into the new log file. In this way we can keep track of all previous events, even the ones created by the producer of the SIP. Corresponding import will happen when a check-in to the control area is executed.

Check-in to control area

When IP's in the work area are to be checked-in to the control area they are actually moved into the control area. A check-in to control area is activated by an authorized logged in user in EPP. A list of possible IP's to check-in can be found in the users work area. A selection of IP's has to be done in order to be checked-in. When selected processed IP's are checked-in they are moved from users work area into associated AIC directory in the control area. A new content description *mets.xml* will be created due to the new structure of the processed IP. The log file *log.xml* will also be imported into EPP database during check-in. After the check-in we do have the original IP and the processed IP in the same parent AIC directory.



AIC_UUID – unique directory for several IP's related to each other
IP_UUID – unique directory for original IP_0
IP_UUID – unique directory for processed IP_1

10. Work area

The work area is used by authorized personnel within the preservation organization, for different processing of IP's. Processing can be complements, updating or adding metadata, format converting, removal of unwanted files etc. Processing events are manually registered by ET into the log file *log.xml*.

A newly checked-out IP is built upon the directory structure used when the SIP where received in the blue zone. An original IP can be processed in many ways. It is appropriate to use the same directory structure and to complement its content. Checked-out IP's are considered to be processed once they are checked-in again.

Sometimes it's preferred to add some information not contained in the SIP. This can be done after receiving the SIP and be a part of the archival processing in the work area. If, for example, the SIP is found to be incomplete in some way, it could be addressed to the producer as a request to resend the SIP or addressed as complementary request for information.

The data-lock (exchange area) is being used for external information exchange from and to the control area. It is only by EPP you are authorized to exchange information between the secured orange/red zone and the blue zone. The exchange area is monitored closely and continuously and all transactions are being logged.

If a request of exchange is initiated from within EPP it must be addressed and detailed as a 'query request' or 'update request'. The request type needs to be interpreted and checked for periodically. Along with the request there could be attached objects (files), for example if a request for complementary information, such as EAD[4]/EAC[5], is sent to the exchange area by EPP, it should be interpreted by the AIS and a query result set should be returned to the exchange area, in order for EPP to get the result set from the exchange area. The other way around could be when the AIS needs to be updated with newly checked-in SIP's and the AIC information.

Any information can be exchanged with the exchange area as long as it is initiated by EPP. Therefore it is important for EPP to continuously check for any request types at the exchange area and to be prepared to interpret it and to act according to that.

11. Check differences

Every IP has its own content description *mets.xml*. A so called 'Diff Check' is activated by an authorized logged in user in EPP. A list of possible IP's to diff-check can be found in the control area. A selection of IP's has to be done in order to let them be diff-checked. When selected IP's are diff-checked they are compared with their associated original IP described by AIC. It is the content description *mets.xml* of the original IP that is compared with the physical content of the selected IP.

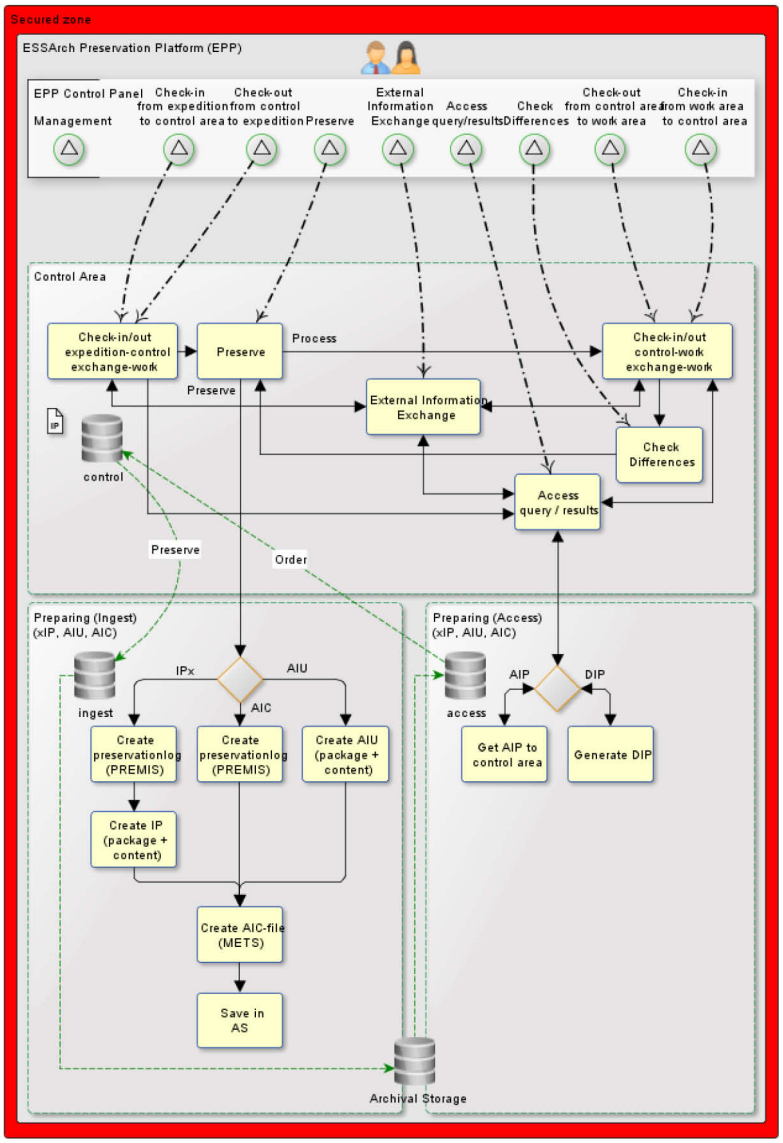
Differences and deviations discovered regarding added, deleted or updated files are presented to the user in the EPP GUI. If it turns out that there are differences or deviations, they need to be taken care of, especially if selected IP is to be approved for archiving. If necessary, an IP can be checked-out again to work area and be further on processed, checked-in again and diff-checked once again. You can at any time activate a diff-check within the control area.

[4] EAD - Encoded Archival Description is an XML standard for encoding archival finding aids, maintained by the Technical Subcommittee for Encoded Archival Description of the Society of American Archivists, in partnership with the Library of Congress.

[5] EAC - Encoded Archival Context is an XML standard for encoding information about the creators of archival materials and the circumstances of record creation and use. It can be used in conjunction with EAD for enhancement of EAD's capabilities in encoding finding aids, but can also be used in conjunction with other standards or for standalone authority file encoding.

5.2 Scenario – Delivery from the e-archive

An access request can be compiled in many different ways. Even the consumers have to be clarified and explained, as well as credentials. But one thing for sure is that if archived information is not available and accessible it is not usable to anyone. The whole idea of e-archiving is lost in that case. So, of course, EPP does have capabilities of serving those types of requests.



Access requests must be interpreted and validated before they are executed. The questioner can be a consumer or a producer or another instance within EPP. To be able to retrieve any information from EPP one has to be authorized to do so. It is controlled by EPP.

Access requests can be initiated by archival processing in the work area as well as from external resources and procedures in the blue and white zone.

An EPP external access request must be checked-in as query/result request in order to be able to execute corresponding actions.

The result of the query can either be from an order, eq. DIP or an AIP retrieval, or a result set of metadata.

The result from a query can be delivered to either the work area or the exchange area.

First a request type has to be defined. The purpose of the request can be clarifying why, when and how often it has to be executed. The result can be used in different ways. To be able to make an access request, one has to be authorized by EPP. By this, a potentially questioner already are authorized to put together a query of any kind.

The request is validated and authorization controls are performed, for example, if the query is compiled correctly, if user is authorized to execute the query or if only the result set is restricted to user etc. Many different validation tasks are done but the most important ones are the ones who check the compilation and authorization.

If a request type is compiled and interpreted as a search query the task is performed and the result set is returned from the assembling functionality to the questioner who is authorized to receive the result set.

If the request type is not a search query it has to be of another type and therefore accurate preparations are made according to interpreted request type. For example, if the request type is an order, preparations are made to search for and to retrieve an archived IP in the archival storage. The AIP has to be available and accessible and therefore the query must be passed on to access functionality within the I/O engine of EPP.

Before the query is passed on to the I/O engine it needs to be validated and if necessary restructured. A query has to be compiled correctly and express something that EPP can understand and answer to. All these validations are done in order to qualify the query to be passed on to the I/O engine.

If an order is queried an AIP is retrieved from archival storage by I/O engine. Based on the form of the order the selected AIP will be retrieved and handed over for further validation.

The result of the query is made available and the interpreted query shows where the result will be delivered. Often the AIP will be made available to the control area in EPP for further processing. The AIP can there from be copied to work area and then to the exchange area.

To check-out an AIP from the control area to the work area does not differ from the procedures when an IP_X is to be checked-out. Due to what the AIP is intended to be used for, it could either be checked-out as a read-only copy or as a new IP. If it will be checked-out as a new IP it will be considered to be prepared for a check-in to the control area again. A read-only copy is not possible to check-in again to the control area.

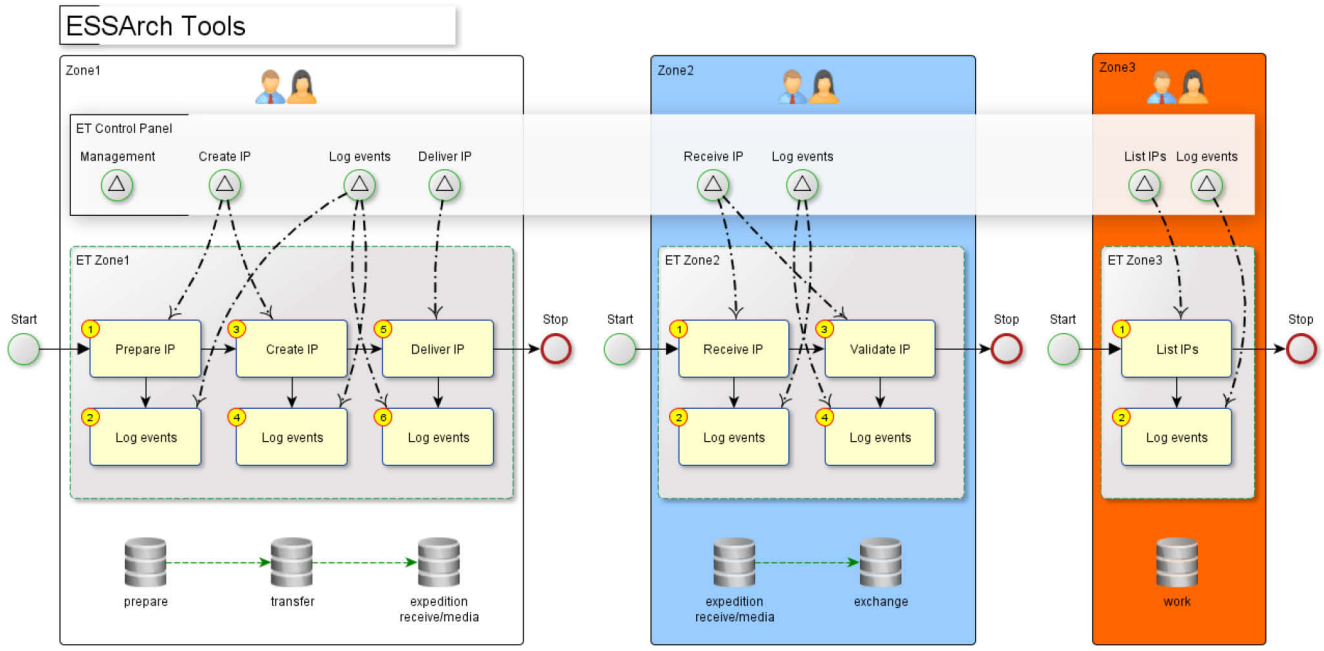
Different processing can take place in the work area. One of the most common ones for a retrieved AIP is to convert or restructure it into a DIP eq. AIP-DIP conversion. The DIP can then be checked-in to the control area again or checked-out to the directory used for these types of requests at the exchange area.

6. ARCHITECTURE OF ESSARCH

6.1 ESSArch Tools (ET)

6.1.1 System Architecture

ET is an application (software) developed in the Python language and the web framework Django. The application is constantly evolving and is aiming to be platform independent and is licensed as Open Source.



ET can be profiled as three different zones and installed as the chosen profile. A producer is represented by zone1, a receiver in a preservation organization by zone2 and zone3 represents the secured environment where the EPP could be installed or another digital e-archive (TRAC).

Each zone has its own configuration parameters and paths to directories. ET can be used with local directories or network addressed file systems. Normally, after a basic installation, ET will be installed as a producer with all required configuration profiled as the country Norway. The country specific configuration as well as paths can easily be changed within ET if logged in as an administrator.

6.1.2 Code and directories

Description of directories and code.

ESSArch Tools – ESSArch Tools default standard path: /ESSArch/Tools

Program – Default standard path: /ESSArch

Logs – ESSArch Tools system- and debug logs. Default standard path: /ESSArch/log

Configuration – environment and mimetypes. Default standard path: /ESSArch/env

Directory structure for zone1. All paths are initially /ESSArch/SIP_room/<path>

Prepare – information package is prepared before creation and delivery

Transfer – information package is created and ready for delivery eq. removed from prepare path

Expedition – the receivers dedicated area for receiving deliverable SIP's

Directory structure for zone2. All paths are initially /ESSArch/expedition/<path>

Receive – receivers dedicated area for receiving delivered SIP's

Exchange – data-lock for external information exchange between blue and orange zone.

Media – reception of the preservation organization and area for SIP's to be checked-in to control area

Directory structure for zone3. All paths are initially /ESSArch/<path>

Work – work area is used for test and processing of received IP's by the preservation organization

6.2 ESSArch Preservation Platform (EPP)

6.2.1 System Architecture

EPP is implemented in the red zone, the secured zone, and EPP is within that zone the only application with permission to access resources such as file systems, archival storage etc. EPP also exchanges information with external resources such as AIS through the exchange area. This area (exchange) is the only way to exchange information between orange/red and blue zone and it is only EPP who is authorized to do so.

6.2.2 Code and directories

Directory structure for EPP.

EPP – Root directory for EPP. Default standard path: /ESSArch
Application – Binaries and executable scripts for EPP. Default standard path: /ESSArch/bin
Logs – System logs. Default standard path: /ESSArch/log
Debug Log System – If any of EPP system processes should be interrupted without control a debug log will be generated. Default standard path: /ESSArch/log/proc
Debug Log XML validate – If schema validation results in deviations or errors, result information will be stored in this directory. Default standard path:/ESSArch/log/debug
Configuration Base – Configuration files for EPP, Apache etc. Default standard path: /ESSArch/config
Configuration Application – Configuration files for EPP GUI etc. Default standard path: /ESSArch/app/config
Expedition – Area used for receiving SIP's that will be checked-in to the control area. Default standard path: /ESSArch/media
Control area – Area only accessible by EPP, used for transactions. Default standard path: /ESSArch/control
IngestPath – Area used for IP's which are to be archived. EPP automatically detects and verifies the quality of the IP before it is archived. Default standard path: /ESSArch/control/ingest
AccessPath – Area used for DIP's. Default standard path: /ESSArch/control
Work Area – Area used for test and processing of IP's. Default standard path: /ESSArch/work
Exchange – data-lock for external information exchange between blue and orange zone. Default standard path: /ESSArch/exchange
EPP DB backup – EPP schedules backups of local database to a backup file. Default standard path: /ESSArch/backups_mysql
Library – EPP lib-modules. Default standard path: /ESSArch/pd