Research and Practice on SIP Ingestion Based on Trusted Workflow Management

Wu Zhenxin
National Science Library, Chinese Academy of Science
BeiSiHuanXiLu 33, Beijing 100190
86-10-82629453, China
wuzx@mail.las.ac.cn

Liu Jianhua
National Science Library, Chinese Academy of Science
BeiSiHuanXiLu 33, Beijing 100190
86-10-82629453, China
liujh@mail.las.ac.cn

Gao Jianxiu
National Science Library, Chinese Academy of Science
BeiSiHuanXiLu 33, Beijing 100190
86-10-82629453, China
gaojianxiu@mail.las.ac.cn

ABSTRACT
From the perspective of trusted workflow management,this paper discusses research and practice on ingest management of digital preservation system of electronic journals. It first describes the trusted workflow management model and trusted chain mechanism, and then strategies on data package management and assembly workflow construction are described in detail. In addition, it divided the ingesting workflow into atomic processes combining with the actual processing requirements, and made a personalized workflow definition and processing demonstration taking IOP for example.

Keywords
Trustworthiness Workflow Digital Preservation Ingestion Management

1. INTRODUCTION
Because of factitious mistake, technical upgrading, equipment damage and other reasons, it often results in a continuing decay and loss of integrity, authenticity, security and usability of digital objects, which is an important issue we must face in research and practice of digital preservation. As the digital preservation system play an important role in digital preservation, it need make use of a variety of strategies, technologies and methods to keep integrity, authenticity, security and availability of data objects.

In a digital preservation system, data ingesting module is the initial entrance of all digital objects which will be archived. It plays as a bridge for information transfer between digital preservation system and content providers. From receiving the information package (SIP), it carries out a series of related processes and finally creates an effective Archival Information Package (AIP) complying with archiving data format and data standards. So, effectively control on the ingesting processes of the original SIP will directly affect the quality of the data archived in the system, and the ingesting module is the first step to ensure integrity, authenticity, security and availability of archiving resources.

There are already some digital preservation systems doing research and practice on ingesting management based on different context and demand, such as the e-Depot and Portic, which formed the distinctive ingesting management functions and workflows. Based on the same purpose, we did in-depth studies on the trusted workflow management during developing our digital preservation of electronic journals.

2. TRUSTED WORKFLOW MANAGEMENT MODEL AND TRUSTED CHAIN MECHANISM
Digital preservation is a complex systems engineering. There are some differences in requirement on data control and management between itself and other information system. Besides, it is difficult to find and correct the mistake during preservation process in short time because of the specialty of digital preservation. Therefore, it need stricter workflow management and control for digital preservation system.

Take into account related requirement of process management and trusted archive authentication, we proposed the trusted workflow management model (figure 1).

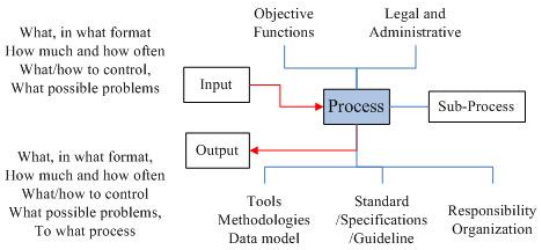

Figure 1. Trusted workflow management model[1]

According to the trusted workflow management model, we should define some information for each course as follows:

（1）Atomic process definition and basic requirement

We need define objective of each atomic process. In other word, we should confirm the operation tasks and its functions, referred technical specialty, performance requirement, related law limitation and management requirement and so forth.

（2）Input information

Meanwhile, we should be conscious of the input requirements of each atomic process which include the type, format, amount, input frequencies of input information. Besides, we should know

how to control the input information and how to deal with the problems during the information import.

（3）Output information

Similar to the input information, here we should demonstrate the type, format, amount, output frequencies of output information, and identify how to control the input information and how to deal with the problems during output.

（4）Process

Information management process is the core of workflow. Any activity that converts input resource into output one will be regard as a process. Each process could contain multi sub-process and the output of previous process might be input one of next process. To insure the efficiency of digital preservation, the system should identify and manage many related and interactive process. There are usually four elements in a digital preservation trusted workflow.

- Information. It refers to the related data resources such as inner information, exterior information and flow control information. All of these are used to describe process of workflow and expressed as digital preservation policies, procedures, guidelines and so on.
- Method. It contains standards, technologies and some methods for support other resources which would be used in digital preservation.
- Organization and responsibility. This element describes each entity and their relationship within workflow process. It is represented as digital preservation mechanism, personnel requirements, and work report systems and so on.
- Activity. Activity represents each process, sub-process and their restrict relationships which form a workflow. All these activities will turn into a complete workflow through some control manners such as ranking, combination, parallel, serial, repeated.

The Trustworthiness of a workflow is reflected in its own scientific, reasonable, trusted design. On the other hand, an obvious, clear, open and verifiable description or prescript about process and control practices of the workflow can also enhance its trustworthiness. From the perspective of process management, it requires related criterion, standards and management systems to implement the workflow. But also it need use related criterion, standards and management systems to insure trusted management of workflow. Therefore, we could make use of the inspection of some criterion, standards and management systems which are indispensable to evaluate the trustworthiness of a workflow.

To trusted chain mechanism, it means that a certain task is divided into workflow chain consist of successive, multi atomic process. The trustworthiness of each atomic process is based on trustworthiness of process context and previous process function of system. So, we could guarantee the trustworthiness of each atomic process though strict control management and insure trustworthiness of the whole flow via constructing a trusted chain of workflow.

3. SIP INGESTION MANAGEMENT BASED ON TRUSTED WORKFLOW

The digital preservation system (DPS) of national science library applies Fedora as the substructure core repository. Considering the workflow complying with Open Archival Information Service reference model (OAIS), the requirement of trusted repository and actual demand on preservation, DPS provide a series of preprocesses on SIPs to support the next archive management.

3.1 Strategy on Data Package Management of DPS

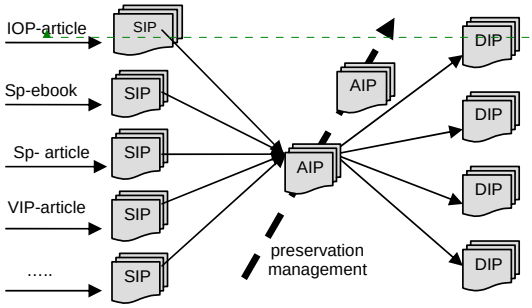

Figure 2. Strategy on data package management of DPS

Currently, SIPs are from different suppliers. They can't be submitted in the light of a uniform standard format. In this case, the DPS adopted a strategy in ingesting module design. It "receiving SIP in different formats, submitting AIP in uniform formats, distribute DIP in different formats". In other word, It allows the system to receive and process SIP in a variety of formats, and then generate a unified format of each SIP for archiving management.

3.2 Strategy on Assembly Workflow Construction

Before being ingested into archiving system, SIP in different format need go through various preprocess. Therefore, the ingesting system must be able to provide a more flexible workflow construction strategy, and offer customized workflow management for different submission format. According to the modularization program development thinking, the DPS divided ingesting process into many atomic processes. And then it defines the atomic processes one by one according to the trusted workflow management model, and develops modules separately for each atomic process. In the process of ingestion, operators can choose required atomic processes in term of the preprocess demands of SIP in different formats. They may add personalized information (such as documents, tools, standards and responsibility individuals), config and sort these atomic processes to form a personalized workflow.

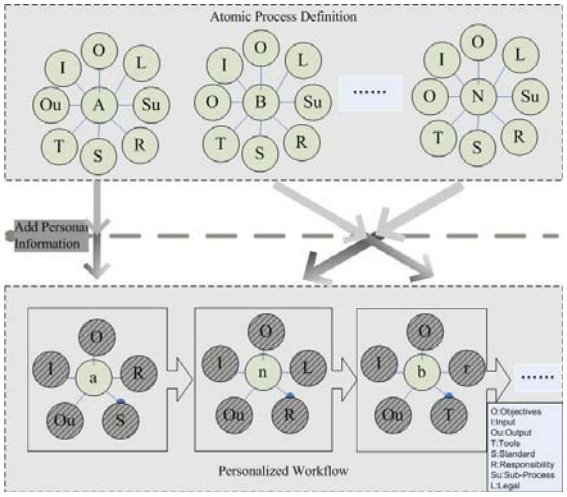

Figure 3. Assembly workflow constructions

3.3 Ingesting Workflow Decompose

There are some descriptions of ingesting module in the OAIS standard. But the OAIS model is only a conceptual one. We need refine the steps which don't have detailed definition according to our own demands in practical preservation system, such as data auditing, responsibility allocation, data semantic definition, workflow model standard and so on.

In addition, most people proposed some corresponding requirements in ingestion phrase in the trusted study of preservation repository. For example, the Nestor criteria catalogue claims that: Repository should define relative specifications of SIP from suppliers to ensure the integrity of digital object; identify the risk of digital objects migration; ensure safe transmission from supplier to repository; ensure integrity and quality of transmission. Criteria standards in OCLC' official release of "Trustworthy Repositories Audit &Certification: Criteria and Checklist (TRAC)" require that ingesting module should provide safeguards of digital objects' source, correctness, integrity and full control.

Based on these studies, we divide ingestion management into 10 detailed steps according to practical data ingestion and process: SIP Receipt, Transmission Integrality Check, Virus Check, Unzip, SIP Count Check, SIP Format Check, Metadata Check, Standard SIP Formation, Standard SIP Check, Archive. Then, in accordance with the trusted workflow model, we give detailed description of documents, tools, stuff, processing specification and other things required by each atomic process.

(1) SIP Receipt: Receive data and related documentations from suppliers; carry out initial registration of this batch of data according to the documentations, including data sources, data type, the given time, the receipt time, the recipient, the time of archiving, and so on.

(2) Transmission Integrality Check：Use the Checksum of original SIP for data integrity check.

(3) Virus Check：Detect virus and Trojan.

(4) Unzip: Unzip the archives to specified directory by the rules.

(5) SIP Count Check: Count the numbers of various documents, check the path and relationships of them and compare the check result with the checklist submitted with the package by suppliers.

(6) SIP Format Check: Check the formats of XML and PDF of initial submitted SIP.

(7) Metadata Check: Check the fields and content of metadata using pre-defined XML structure and content.

(8) Standard SIP Formation: If the package is not a standard SIP, it will be generated into standard one. Meanwhile, extract related metadata.

(9) Standard SIP Check: Check the standard SIP before uploading.

(10) Archive: Submit the standard SIP into the preservation system for archive.

4. CASE STUDY ON DATA INGESTION MANAGEMENT

In our digital preservation system, we define atomic processes (including basic functional description, input and output information, related standards, criterion and technical methods, etc.) in the atomic processes management module of system management as flows (figure 4).

Figure 4. Web page for atomic process definition

In the process management module, we can define a custom workflow for each resource which will be ingested. Fig. 5 shows how to define a workflow for IOP. Firstly we select atomic processes which we need, then append personalized information (related requirements and responsibilities of staff, related policies, documents,

manuals, work guide) of each atomic processes, sort them in need, finally form a personalized ingesting workflow.

Figure 5. Web page for custom workflow definition

During the ingesting, we will choose a pre-defined workflow for each package, after that, the system will call the atomic processes according to workflow. At the same time the system will provide the relevant information which is appended during workflow definition for the operator at the suitable time. After the process of each atomic process,the system will give recommendations for processing result. At the end of the entire process, the processing report and results will be generated. If any problem appears in any atomic process, the data package will be shifted into the error management and wait for manual handling.

Our DPS provides two kinds of processing approach: manual one and automated one while the automated one hasn't been completed.

5. Conclusions

After a lot of testes on some kinds of data packages, our design of ingesting workflow management was verified to be appropriate. It basically met with our requirements for flexible, customizable, personalized and scalable of the workflow management besides responding ingestion operation.

Ingestion processing of digital preservation system is actually performed by a coherent set of processing steps (atomic processes) with cooperation. The data packages flows between the different processes in accordance with pre-defined workflow, completes processing on different kinds of digital resources with the detailed specifications and system requirements. The ingesting workflow management program discussed in this article, division entire workflow into a series atomic process, defines functions and requirements of each step particularly, and lists specific standards and tools are used, and ensure integrity and availability of digital objects in the ingesting process, provides trusted support for the follow-up archive management. Related documents, recommendations and a detailed record of the process, make ingestion management has a very good transparency and intelligibility. As a complex application system, digital preservation system should have the trusted characteristics, the trusted ingesting workflow management make a good foundation for the trusted digital preservation system.

6. REFERENCES

[1] Li Chunwang, Zhang Xiaolin, etc. NSTL research report on trusted workflow management of digital preservation system, 2007.