The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0956-5698.htm

# Algorithmic methods to explore the automation of the appraisal of structured and unstructured digital data

Basma Makhlouf Shabou
*Geneva School of Business Administration, HESSO, Geneva, Switzerland*

Julien Tièche
*City Archives of Montreux, Montreux, Switzerland, and*

Julien Knafou and Arnaud Gaudinat
*Geneva School of Business administration, HESSO Geneva, Geneva, Switzerland*

Received 1 September 2019
Revised 6 February 2020
Accepted 2 March 2020

## Abstract

**Purpose** – This paper aims to describe an interdisciplinary and innovative research conducted in Switzerland, at the Geneva School of Business Administration HES-SO and supported by the State Archives of Neuchâtel (Office des archives de l'État de Neuchâtel, OAEN). The problem to be addressed is one of the most classical ones: how to extract and discriminate relevant data in a huge amount of diversified and complex data record formats and contents. The goal of this study is to provide a framework and a proof of concept for a software that helps taking defensible decisions on the retention and disposal of records and data proposed to the OAEN. For this purpose, the authors designed two axes: the archival axis, to propose archival metrics for the appraisal of structured and unstructured data, and the data mining axis to propose algorithmic methods as complementary or/and additional metrics for the appraisal process.

**Design/methodology/approach** – Based on two axes, this exploratory study designs and tests the feasibility of archival metrics that are paired to data mining metrics, to advance, as much as possible, the digital appraisal process in a systematic or even automatic way. Under Axis 1, the authors have initiated three steps: first, the design of a conceptual framework to records data appraisal with a detailed three-dimensional approach (trustworthiness, exploitability, representativeness). In addition, the authors defined the main principles and postulates to guide the operationalization of the conceptual dimensions. Second, the operationalization proposed metrics expressed in terms of variables supported by a quantitative method for their measurement and scoring. Third, the authors shared this conceptual framework proposing the dimensions and operationalized variables (metrics) with experienced professionals to validate them. The expert's feedback finally gave the authors an idea on: the relevance and the feasibility of these metrics. Those two aspects may demonstrate the acceptability of such method in a real-life archival practice. In parallel, Axis

© Basma Makhlouf Shabou, Julien Tieche, Julien Knafou and Arnaud Gaudinat. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial & non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at: http://creativecommons.org/licences/by/4.0/legalcode

The authors' gratitude goes to their institutional partners Lionel Bartolini, Pascal Schneiter and Grégoire Oguey from Neuchâtel State Archives and Matthieu Desjobert from Telono with whom they had the chance to experience their ideas. Their recognition and thanks also go to Kamel Nebhi and Aurèle Nicolet from HEG for their invaluable help in editing this work.

Records Management Journal
Emerald Publishing Limited
0956-5698
DOI 10.1108/RMJ-09-2019-0049

2 proposes functionalities to cover not only macro analysis for data but also the algorithmic methods to enable the computation of digital archival and data mining metrics. Based on that, three use cases were proposed to imagine plausible and illustrative scenarios for the application of such a solution.

**Findings** – The main results demonstrate the feasibility of measuring the value of data and records with a reproducible method. More specifically, for Axis 1, the authors applied the metrics in a flexible and modular way. The authors defined also the main principles needed to enable computational scoring method. The results obtained through the expert's consultation on the relevance of 42 metrics indicate an acceptance rate above 80%. In addition, the results show that 60% of all metrics can be automated. Regarding Axis 2, 33 functionalities were developed and proposed under six main types: macro analysis, microanalysis, statistics, retrieval, administration and, finally, the decision modeling and machine learning. The relevance of metrics and functionalities is based on the theoretical validity and computational character of their method. These results are largely satisfactory and promising.

**Originality/value** – This study offers a valuable aid to improve the validity and performance of archival appraisal processes and decision-making. Transferability and applicability of these archival and data mining metrics could be considered for other types of data. An adaptation of this method and its metrics could be tested on research data, medical data or banking data.

**Keywords** Archival appraisal, Algorithmic method, Appraisal criteria, Appraisal metrics, Automation appraisal, Data mining

**Paper type** Research paper

## Introduction

This study presents the results of a practical and theoretical one-year research guided by the needs of the Neuchâtel State Archives in Switzerland (OAEN) and led by a team of researchers in the Information Sciences Department at the Geneva School of Business Administration (HESSO Geneva). The challenge proposed by the Neuchâtel Archives was to be able to handle an extreme case of data appraisal where the archivist would have to manage, for example, several digital media (hard disk or optical media) without any indication of their nature and their context. To meet this challenge, the idea explored was to make maximum use of data mining and artificial intelligence approaches to facilitate and prepare the archivist's appraisal phase. The objective was to identify possible functionalities to be introduced in appraisal software, to help define the specifications of such a tool. The challenge was also to take into account all possible and available structured information in a variety of scenarios, ranging from creating organizations with low-record management maturity, through to organizations with well-developed records management maturity. On top of this approach, a model of archival metrics, as proposed by research on appraisal criteria and their metrics (Makhlouf Shabou, 2011a, 2011b, 2015a, 2015b), was studied to consider data mining in a business framework.

## Context and problem

The overarching objective of this research was to develop a proof of concept for an archival appraisal tool that will assist in decision-making regarding the archiving or disposal of structured or/and unstructured corporate data sets while having a valid and defensible argument and method. Under the current archival practice and the law on archiving in force in the State of Neuchâtel, organizational units in the state propose their closed files or unstructured data at the end of their useful life as archives to OAEN. OAEN, the authorized authority by delegation of the Council of State, evaluates and applies the appropriate records disposition to the proposed closed files or data packages. The proposal takes place at the end of the records' administrative or legal utility, and the proposal is the responsibility of the records management officer, appointed within the organizational unit (Loi sur l'archivage, 2011).

At the OEAN, in this project, the proposed transfer of closed files or data packages could be carried out according to the two scenarios that the appraisal software tool would need to deal with:

> *Scenario 1:* Enclosed files or data packages are provided in the appraisal software tool, with a structure, retention schedule, organic links, context and description sufficiently developed. The intended final life cycle period and records disposition are clearly identified according to a previous prospective appraisal, conducted jointly by the archival entity and the organizational entity (archives producer).
>
> *Scenario 2:* The data and digital objects are received without any prior documentary or archival processing and are received as unstructured data. The expected tool accommodates the data to be proposed and briefly analyzes and produces a general picture to identify closed files or data sets.

The appraisal software tool would need to accommodate a prospective evaluation by an archivist, but would also need to apply a set of general and specific criteria as well as quantitative and qualitative, intrinsic and extrinsic metrics automatically, semi-automatically and/or manually.

The main potential users of this proposed appraisal tool would be archivists who are required to conduct a rigorous appraisal process in a documented way to reach a defensible decision.

The proof of concept project to develop the proposed functionality and metrics that would drive the appraisal software tool is explored through the remainder of this paper. The literature review that follows identifies a range of sources that informed the development of metrics for both archival value and data mining in an appraisal context. The methodology section explores how different methods were used to develop metrics to calculate an "archival value" rating as well as data mining insights. The findings section demonstrates the outcomes and learning that were achieved through this work.

## *Literature review*

Before discussing the work done in the study, the following fundamentals of archives appraisal influenced the work. Appraisal is a major archival function that manages the life cycle of records. The objectives of appraisal are to determine which documents to create and capture, how long different corporate records need to be kept and to identify their final disposition. An appraisal clarifies and documents responsibilities and roles over a record's lifespan. It is operated by a documented process and tools that are at once collaborative and should help the decision-making process. This process considers different internal and external contextual dimensions (International Organization for Standardization, 2018; Makhlouf Shabou, 2019). An appraisal is a very challenging process, and in the digital era, automation opportunities can be incorporated into appraisal processes to support human decisions and actions.

### *Archival metrics*

Recent European projects, Australian and North American initiatives have focused on the automation of appraisal functionalities, and these projects have influenced the design and functionality of the proposed OEAN archival appraisal tool.

In Germany, at the Ludwigsburg State Archives, a "More Products Less Process" (MPLP) approach was adopted between 2016 and 2018 for the appraisal of a large digital collection (Belovari, 2018, 2019). The project tested ten deduplication software packages (80% failed). It selected a user-friendly and inexpensive software package. The project then used this software to assess 670 GB of data in four days. Based on this, an adapted workflow was developed to support appraisal. This initiative confirmed the potential benefits that may be expected using software to support appraisal processes. As Belovari (2018) states:

> In our project collection, 15% of Word documents but only 5.6% of graphic and video files were duplicates. This kind of information may feed into records management and donor conversations and may improve how files are kept and what is transferred.

In the UK, the MPLP approach is being considered at the University of Westminster. As part of the Research Data Shared Service Digital Preservation Pilot Program, Penn (2019) identifies how the university is exploring the archival and records management aspects of digital preservation. The program has hosted a workshop for UK practitioners on the topic of appraisal "where we explored how well traditional appraisal theory and practice can be applied to digital records and how we document these processes" (Penn, 2019). A conclusion of the research to date is that the MLMP method remains useful for appraising digital records.

In a public French administrative entity, a software package dedicated to electronic file plans has been developed. Archifiltre (https://archifiltre.github.io/) is software jointly developed through the collaboration of a data scientist, an archivist and an IT developer. It has been adopted by the social ministries in the French government. This tool enables the acquisition and description of digital structured contents. It offers a qualitative and quantitative mapping of contents to facilitate their visualization and selection. It is made available to public archives producers who wish to transfer their documents to the national archives. While transfer is not automated, Archifiltre visualizes data, detects copies and content redundancy and enables metadata completion and chronological description.

In Canada, other researchers have explored how machine learning and assisted classification may facilitate the appraisal of emails as records of value for the organization (Vellino and Alberts, 2016). The researchers report that:

> The study performed a qualitative analysis of the appraisal behaviours of eight records management experts to train a series of support vector machine classifiers to replicate the decision process for identifying e-mails of business value. Automatic classification experiments were performed on a corpus of 846 e-mails from two of these experts' mailboxes (Vellino and Alberts, 2016).

This study shows that the need to automate e-mail processing (classification and appraisal) is becoming obvious, and the growing amount of content requires more efficient methods to ensure appropriate recordkeeping processes and decisions (conservation and/or deletion). Business value is the main criterion considered in this study, and some metrics for assessing business value were proposed. However, their application is dependent on previous content tagging, which, in most cases, is a time-consuming and manual task.

On the use of artificial intelligence approaches for archives, Rolan *et al.* (2019) provide a good overview of the literature from expert systems to deep learning models. They cite four concrete examples in Australia, including a proof of concept (PoC) of machine-assisted email appraisal by the Public Record Office Victoria, an automatic classification pilot for appraisal and disposal by the New South Wales State Archives and Records, a research project on using text mining for disposal by the National Archives of Australia and finally, use of microservices architecture and linked technologies for automating record management by the Australian government Department of Finance.

Considering the emergence of computer-assisted appraisal for records management, Harvey and Thompson (2010) give a good introduction to assumptions about automation for appraisal. Lee (2018) provides a list of technology opportunities such as digital forensics tools, natural language processing methods and machine learning algorithms.

In the OAEN project, a systematic review of the literature was used to assess the academic literature that identifies the value and quality of data, documents and archives as well as the professional literature which identifies the different standards relating to information and documentation. In addition, to support the requirements of the Neuchâtel public context, legislative texts relating to information and archives were analyzed.

The analysis of these different sources highlighted the various qualities or characteristics that archives must possess. Indeed, according to the ISO 15489 standard on records management, documents must have "characteristics of authenticity, reliability, integrity and exploitability in order to constitute proof of the events of the activity or successful operations and to fully meet the professional requirements" (International Organization for Standardization, 2016, p. 4). The literature review highlighted various possible attributes pertaining to the authenticity, reliability, integrity and operability of the archives. These attributes became part of the conceptual framework for this project. In total, from the literature, three main dimensions were identified to help measure, within the appraisal tool, the potential value of structured and unstructured data – trustworthiness, exploitability and representativeness.

## Methodology

A qualitative and quantitative approach was used for this study. Given the exploratory character of the study, the qualitative method, in part described above, was used to explore the mechanisms and approaches to value informational sources. The quantitative facet was used to identify algorithmic methods and metrics that may be useful to enable the calculation and the scoring of data, records and files in a systematic and/or automated way.

The study combined a top-down and a bottom-up approach. The bottom-up approach explored real data and aimed to identify relevant data mining methods to assist the archivist in documenting an appraisal (Figure 1). The top-down approach studied the field of archival

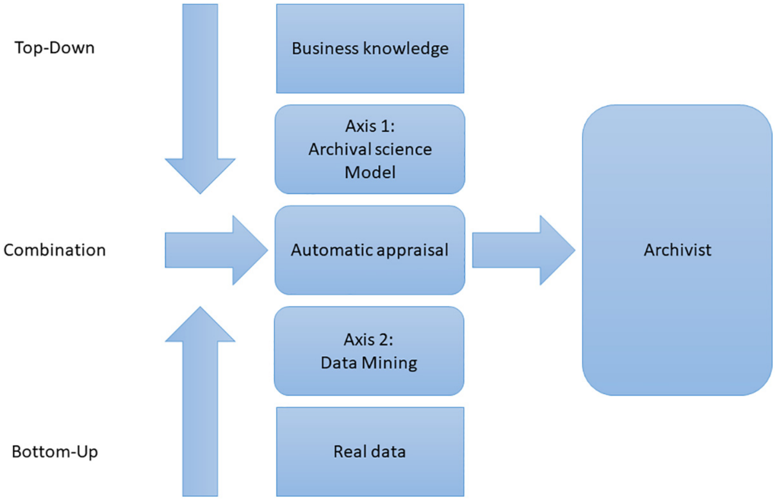


**Figure 1.** Overall methodology, a combination of top-down and bottom-up approach

science to develop a model covering all aspects of archival science. A key challenge of the project is to link these two approaches, i.e. to link archival science metrics with metrics from the data.

The following sections summaries each of these approaches and their findings.

### *Axis 1: archival science data collection*

The mixed character of the approach necessitated mixed data collection methods. A systematic literature review, Likert questionnaire and uses cases were conducted and synchronized between the two main axes of the research.

*Systematic literature review*: a systematic literature review covered the past three decades. A variety of sources were considered in different languages, primarily English and French. Other Arabic and German texts were also consulted. The identified literature included academic projects and studies, best practices and corporate initiatives in public and private entities, as well as standards and normative texts advanced in different European, Asian and North American countries. A predefined reading grid was prepared to guarantee the rigor of this investigation and the relevance of its results. The review generated a set of appraisal metrics.

*Likert test*: The purpose of this step was to evaluate and reach a certain global validation of a set of appraisal metrics by a group of five international experts. These latest contributed also to data mining metrics and functionalities.

Initially, a pre-test was performed to verify the clarity and adequacy of the defined protocol before it was proposed to experts. The choice of experts focused on professionals – archivists or related professionals – with a good knowledge and experience of archival processes, criteria and appraisal tools. A minimum of three experts was required, but ultimately, five experts volunteered to perform the test. Of these, three are active in archives of public institutions (two at the cantonal level and one at the communal level), one person is in charge of private archives in a city library and the other expert works in a private information management and archiving consulting company. The tests were carried out individually to avoid the transfer of bias from one opinion to another. Candidates had to read the available documentation on proposed metrics. The experts were then able to carry out the test by expressing for each measure and variable their level of agreement on a five-point scale:

(1) I totally agree;

(2) I agree;

(3) I neither agree nor disagree;

(4) I disagree; and

(5) I strongly disagree.

They also had the opportunity to make comments or suggestions for improvements. Another test with the same group of experts was conducted on text mining metrics and functionalities.

*Case studies*: Three case studies were developed to concretely demonstrate the nature and relevance of the expected software functionality through archival metrics and data metrics.

The objective was to simulate needs, actions, facilities, roles, actors and processes within the framework of three imagined scenarios representing three levels of records management maturity. These cases were designed to imagine three corporate records management maturity levels to be considered:

*Case 1*: Appraisal of unstructured data sets that are proposed by an entity not particularly well known and whose level of maturity is low and corresponds to Level 1. In this case, the data are imported without previous records processing, meaning they have limited or no description, organization or alignment to good practices and recognized records and archives management national and international standards in force.

*Case 2*: Appraisal of files proposed by an identifiable producer whose records management maturity is at Level 2. In this case, the data are structured in files and/or series, but their records processing is partial and incomplete.

*Case 3*: Appraisal of files of an identifiable producer whose maturity is high and corresponds to Level 3. At this level, the data are imported to the expected appraisal tool. Records processing has been undertaken using established good practices and standards, and the records have been well managed across their entire life cycle.

### *Axis 2: Data mining approaches*

*Data-driven approach with real data (bottom-up level).* The objective of this step was to explore the real data in terms of diversity, volumes, granularity and processing possibilities. The focus was on real data, which an archiving service would be confronted with in the worst-case scenario, i.e. without structured or descriptive information. One exploration was carried out using the raw data from an active network disk shared by the employees of the Neuchâtel State Archives (which we will call dataOAEN). DataOAEN is structured and contains mainly administrative records essential for the activity of the service. To obtain detailed information, data ingestion was performed via the Solr indexing tool that used the Tika content extraction system. Tika is an open-source tool that can detect, extract metadata and text from hundreds of different file formats, including the most popular formats such as Word, PowerPoint, PDF and Excel. Thus, it is possible to browse the tree structure of a hard disk in a comprehensive way, identify records, extract metadata, convert content to full text and index it in a NoSQL database. It is possible to produce descriptive statistics on the content of the various media analyzed via simple queries respecting the syntax of Lucene search engines (the library used by the Solr search engine).

Five small PoC experiments were carried out:

(1) Ingestion of data from dataOAEN, accompanied by an automatic analysis of the raw data, including full extraction and full indexing.

(2) Named entity recognition performed on the dataOAEN collection to extract dates, locations, personal names and service names.

(3) A search for related emails on a collection of emails belonging to an employee of OAEN.

(4) Content type recognition for records such as meeting minutes, using machine learning methods based on a manually developed learning corpus. A recurrent neural network with "long short-term memory" (LSTM) cells to allow for greater information persistence was used for the machine learning. The training set was not used in the test set.

(5) Implementation of some combination of archival metrics with real data metrics from the dataOAEN collection.

*Evaluation of metrics and functionalities by experts.* In combination, the literature review, exploration of real data, analysis of structured data and PoCs were used to identify data metrics and functionalities that may be useful to assist archivists with the appraisal task. The project asked the same experts (already described above) to evaluate the types of possible data metrics and functionalities using a five-point Likert scale.

## Findings

The following presents the findings of the research into developing a model for archival science, the research into data mining and then a synthesis of both approaches.

### *Axis 1: Archival science model*

*Appraisal dimensions and metrics configuration.* The first step allowed the highlighting of a conceptual framework, as shown in Figure 2, built on three broad categories – trustworthiness, exploitability, representativeness.

Across these three categories, 42 variables were defined based on the literature review and also based on tested methods and lessons learned in previous comparable studies (Makhlouf Shabou, 2011; Makhlouf Shabou *et al.*, 2013) . These provided relevant reflections and a theoretical framework of the main concepts that constitute the main pillars of archival appraisal, which can subsequently provide a baseline to measure these concepts. This research proposes that these elements allow the measurement of the value of records and data sets. Metrics have been developed to enable flexible, adaptable and automatable assessments of these variables for appraisal purposes during the appraisal process. The metrics are based on the 42 proposed variables related to the appraisal dimensions (ADs).

*Nature and typologies.* It is important to specify how we distinguish between variables and metrics in our project. A metric is the way to measure the fluctuation of a defined variable in a real and given context. A variable is strongly related to its metric but is not identical to it. For example, in the exploitability category, there is a juridical accessibility variable. The research proposes that the availability of information (AD Level 3) may be measured by two variables: V33 Intellectual protection and V34 data protection and privacy. The related metrics as shown below are proposed as a question to enable the determination of a score for each metric (for more information, see metrics scoring section below) (Table 1).

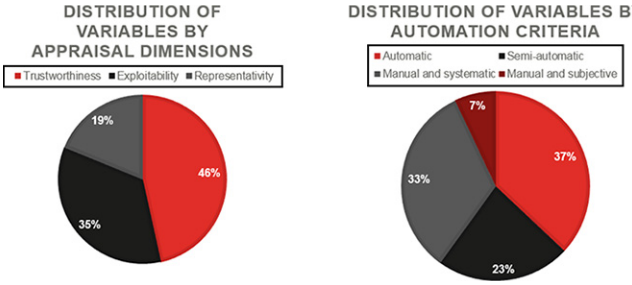


**Figure 2.** ADs and levels

**Table 1.** From dimension to metrics: illustrative example

| | | | |
|---|---|---|---|
| Juridical accessibility | Availability of information | V33. Intellectual protection | *Related metric*<br>Does intellectual property protection restrict the dissemination and exploitation of the record/data? |
| | | V34. Data protection and privacy | *Related metric*<br>Does protection exist that restricts consultation, dissemination and exploitation of the record/data? |

The definition of all 42 metrics is given in Appendix 1. The main typologies of identified metrics may be classified under four main criteria:

(1) *Automation criteria*: Criteria related to how automatable the measurement of each variable is. The measure of the variables can be:
   - automatable: fully applicable by the machine;
   - semi-automatable: partially applicable by the machine;
   - manual and systematic: totally formally applicable by humans; and
   - manual and subjective: totally informally applicable by humans.

(2) *Exclusivity criteria*: Criteria relating to the redundancy of variables' use. The variables can be related to a single AD (exclusive variable, such as V.1 Documentation of transmission) or to several ADs (common variable, such as V.5 metadata completeness, which is used for more than one variable).

(3) *Intrinsic and extrinsic criteria*: Criteria relating to whether the elements used for measurement are internal to the records and data sets (intrinsic), or external (extrinsic) to them, thus referring to its context of creation and/or use.

(4) *Degree of maturity of records management or applicability*: Four levels were defined:
   - no link with maturity levels;
   - *Maturity level 1*: Data sets have not received any processing. These are mainly non processed and unstructured datasets;
   - *Maturity level 2*: The data sets or files have received partial document processing allowing the identification of files and organic series; and
   - *Maturity level 3*: The files have received archival processing on the basis of validated recordkeeping application, including the life cycle management of the records.

### *Metrics measurement: principles and scoring*

The principles underlying the measurement of ADs and their metrics can be summarized as follows:

- The purpose of identifying metrics is essential to support the decision regarding the retention or not of electronic records and data.
- The defined ADs and their metrics are neither exclusive nor exhaustive. They are, therefore extensible and likely to evolve.
- The measurement of the ADs can be applied completely or partially.
- AD measurement takes into consideration the level of maturity of records management in a given context.
- An odd number is considered for the indication of the levels of maturity to be able to identify an intermediate or average level.
- With regard to these levels of maturity too, considering the relative and subjective nature of the value of the files, Level 0 is not used. The lowest level then corresponds to Level 1 (0–25%).
- For each dimension, variables are proposed and structured in three levels of maturity.

- The discriminatory power of the results obtained in the application of the measure greatly contributes to its validity.
- Each variable is measured by one or more metrics.
- The metrics and their method of application are reproducible by several experts.

The relevance of the conceptual framework, including ADs and the resulting variables, is based on adherence to these ten guiding principles, as shown in Table 2.

Regarding metrics scoring, we proposed to express the scoring of variables in terms of percentage.

Three maturity levels are used for calculating the appraisal measurement score: Level 1 corresponds to 25%, Level 2 corresponds to 50% and Level 3 corresponds to 100%. These percentages are given as an indication to convert the measure collected into a quantitative value. Other complex equations may be applied for repartition of those percentages. However, at this stage, we have opted for a simple and computable way for these measures. This enables the calculation of performance level of each metric in a given real situation. Ultimately, the systemization of appraisal scoring will help their automation.

As shown in Figure 3 below, trustworthiness is the richest AD in terms of metrics. Globally, 60% of the identified metrics are automatable or semi-automatable.

### *Expert evaluation and Likert test results*

The results obtained following the consultation of the experts indicate an acceptance rate of the 42 proposed metrics of greater than 80%. The three most accepted metrics were V31 – Content description, V10 – Title existence and V3 – Integration of the record in a document repository/system.

On the basis of the Likert test results, two sets of variables were determined. The first contains potentially automatable metrics (“Automatic” and “Semi-automatic”), while the second includes non-automatable variables (“Manual and systematic” and “Manual and subjective”). A third set of variables has been created for those not finding their place in the first two sets. This set of restricted variables includes variables whose measurement seems difficult at the time of the evaluation. The final choice of appraisal metrics would be customized in the light of corporate needs.

### *Implementation recommendations*

At the end of our mandate, we proposed some recommendations to Neuchâtel State Archives. A phased implementation would be appropriate for the proposed variables. Thus, we recommended prioritizing the implementation of metrics with high automation potential, considering the purpose of this mandate. Secondly, we suggested considering the implementation of variables whose operationalization does not require a high level of records management maturity and which could consequently be applied in any organizational context. In addition, the implementation should take into consideration the metrics that have gained the trust of the experts participating in the Likert test. For example, the implementation might consider developing features that implement the top five archival metrics chosen by the Likert test participants. In addition, implementation could consider those metrics that align to the specific needs of the organization, to prioritize urgent and important issues in a particular corporate context. This confirms the modular and adaptable nature of the proposed software.

**Table 2.**
Appraisal metrics and score application

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| V30. Description of creation context | Does information on creation reasons available? | 0. Unavailable information →Maturity level ❶ | 1. Creation context is known and file plan is used<br>Maturity level ❷ | 2. Creation context is well documented and linked to other documentary materials (guidelines, strategy, laws, rules, etc.)<br>Maturity level ❸ | Exclusive | Semi-automatable | Extrinsic/intrinsic |

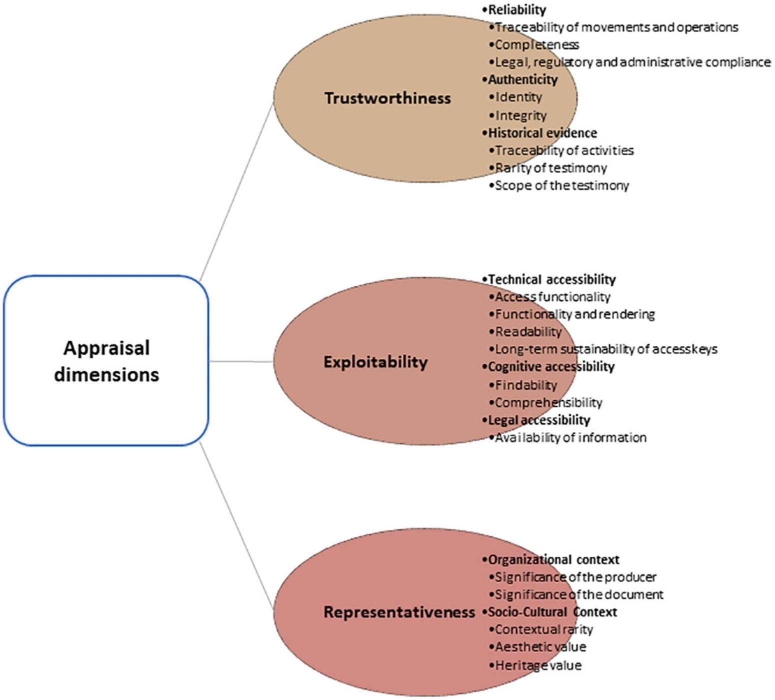


**Figure 3.**
Appraisal metrics distribution by dimension and automation

*Axis 2: Data mining approaches*

*Exploration of real data.* Figures 4–7 give an overview of the dataOAEN collection. This collection has a size of 102 GB for 57,286 documents, where only 13,179 documents have been indexed with metadata and/or full text. Figure 4 shows the distribution of extensions in this collection. For this government department, we observe that there are many image documents (jpg or tiff) that correspond to paper document scans that have not been indexed in the system. The classic office automation files come next (Doc, Pdf, Xls) and correspond mainly to the 13,179 extracted and indexed documents. Note that by observing this distribution, the text recognition functionality becomes absolutely relevant for this service. Without this metric, there is a risk that analysis processes would not have taken into account 77% of the documents.

Figure 5 shows the distribution of indexed documents over time. This can be interesting to identify particular peaks to show the active context (file type, size or organizational metrics). Figure 6 illustrates the diversity of metadata that can be found in the documents. There is at least one document with 350 metadata elements and more than 6,000 documents that have between 50 and 55 metadata elements. Figure 7 shows the distribution of languages within the dataOEN collection. There is a clear predominance of documents in French (fr) because it is the official language of the canton and then documents whose language has not been identified (N/A), English (en) and German (de). Language can be an interesting criterion if we consider, for example, that only official languages of the region should be preserved.

Other information have also been computed and are available in the system but are not shown in this paper (e.g. distribution of file extension for document without date,

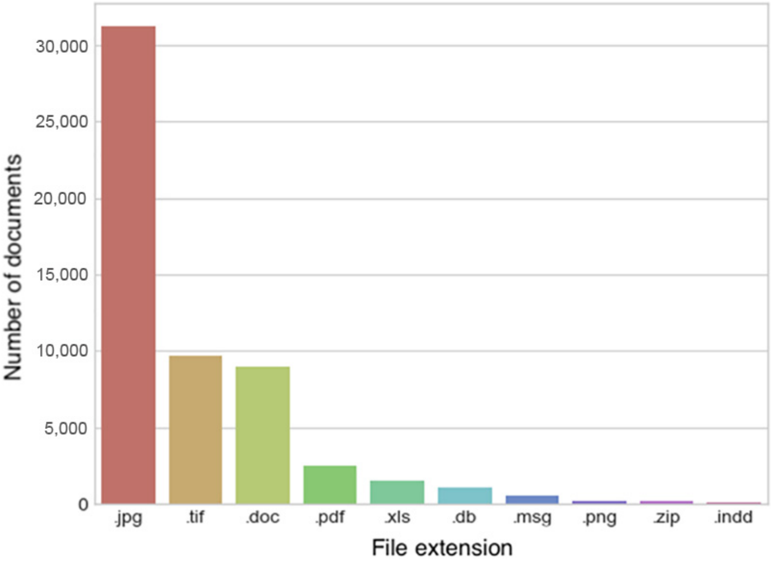


**Figure 4.** Full distribution of the file extension for the dataOAEN collection

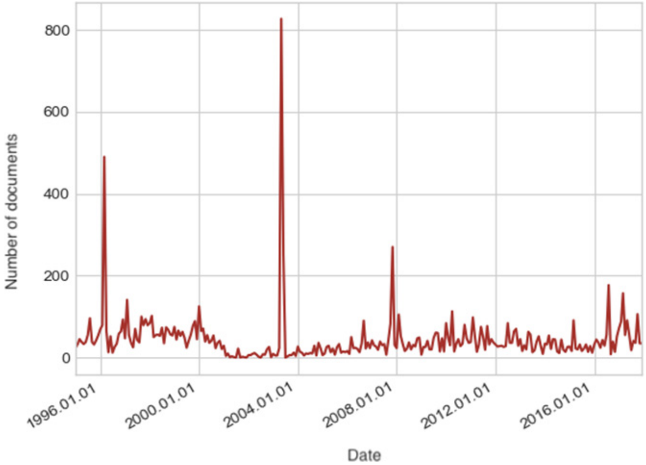


**Figure 5.** Timeline distribution of the number of documents per date for the dataOAEN collection

distribution of file extension for document without author, distribution of file extension for document without content and distribution of some particular metadata). This paper has focused on the information identified as useful to the appraisal process.

*Named entity recognition (NER) and email similarities.* In our PoC, the named entity recognition (NER) functionality was tested for proper name and date recognition. For proper names, we used the "Europeana Newspapers NER" corpora. The names of the people working in the OAEN office were used to verify the approach. For three project partners of OAEN, the number of citations were ranged from 567 to 1,966 for 92 to 755 documents. Then, a word cloud visualization was created to synthetize the results of the proper name

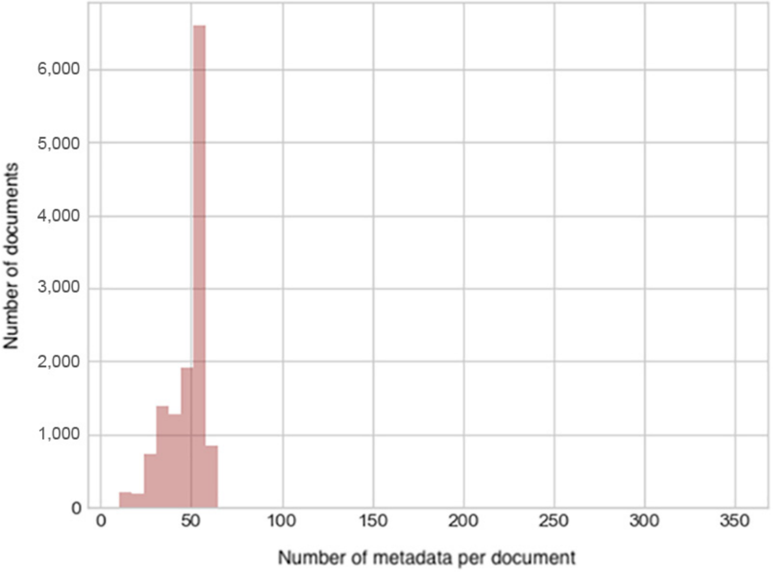


**Figure 6.**
Distribution of number of metadata per documents

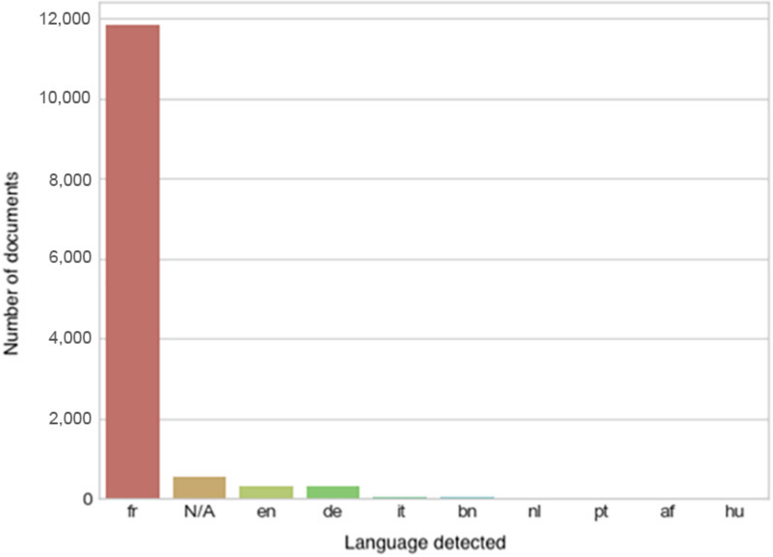


**Figure 7.**
Distribution of the language of documents in the dataOAEN collection

recognition. Thus, the name "Neuchâtel" is the word most emphasized in this visualization because it is most frequently recognized in the dataOEAN collection. Concerning the date, simple regular expressions were used, which allowed us to find 151,429 dates among 7,801 files.

The search for similarity in the emails was done on an additional 2.2GB collection from a PST export (proprietary outlook export format) from a specific OAEN email account. This collection was ingested in the same way as the dataOAEN collection and completed in the same index to make similarity links between email and records documents. A total of 19,358 documents were indexed by Solr, including 10,059 emails, the rest being the attached PDF files. A story line of the documents in this collection was created, and ten similarity searches were carried out for an empirical validation of the approach.

*Automatic classification of topic.* The classification subject used in our PoC was the identification of meeting minute documents. The learning and test sets are presented in Table 3. The test set was created from a Solr query in dataOAEN with keywords related to minute documents in the content and title. The result of the queries was 265 documents. Among them, 210 were identified manually as real meeting minutes.

The system was trained on three completely different collections of dataOAEN. The first came from the shared drive of the Information Science Department and was indexed in the same way as dataOAEN. The same queries as for dataOAEN were made. The second training set came from Google, again with the same queries. The third training set came from a selection of sites belonging to public entities that regularly publish their meeting minutes. The number of documents returned and actually correctly identified as meeting minutes are also given in Table 3 for each training set.

Table 4 gives the performance of the classification system in terms of accuracy. Column 1 shows the percentage of documents from the test set, i.e. dataOAEN, which is also used in the training set. Note that these documents (randomly selected) are no longer used in the test set for evaluation. The last column shows the percentage of minutes documents in the training set that can change as documents are removed for training.

Minutes classification results show an accuracy performance that increases from 88 to 95.8 quickly. This shows that machine learning to assist with appraisal (the topic to be

| Type of set | Query type | From | No. of returned documents | No. of minutes documents (manually checked) |
|---|---|---|---|---|
| Test set | "Minutes" keywords | dataOAEN | 265 | 210 |
| Training set | minutes keywords | Shared drive at information science department | 1,953 | 185 |
| | Minutes keywords | Google | 422 | 318 |
| | Public entities where minutes are published | Selection | 1,035 | 1,008 |

**Table 3.** Design of the test set and the three training sets

| Percentage of dataOAEN test set used in the training set | Accuracy of the system | Percentage of minutes in the test set |
|---|---|---|
| 0 | 88 | 79.2 |
| 5 | 90.5 | 78.1 |
| 10 | 88.7 | 79.1 |
| 15 | 94.2 | 79.2 |
| 20 | 95.8 | 78.8 |

**Table 4.** Accuracy of the classification system depending of test set document inclusion in the training set

defined according to the needs of the application user) can be developed and achieve very good results without too much effort.

*Evaluation of metrics and functionality by archivists.* Following the first steps (study of literature, study of structured data, study of real raw data and realization of PoC), 33 functionalities (see Appendix 2 for full listing) and 72 metrics were identified. From the list of functionalities, six main categories were identified, as described in Table 5.

Both the lists of functionalities and metrics were submitted to our five experts to assess them in terms of agreement and interest using a Likert scale. Table 6 illustrates the overall results with a particular interest for almost 70%, a moderate interest for 24% and little or no interest for 6% of the functionalities.

Table 7 shows the top ten functionalities preferred by the expert. Most of them are related to the fundamentals of search engine and text extraction manipulation, and three are related to the import of raw data and structured data. The optical character recognition one is surprising but, nevertheless logical if we consider that our dataOEAN collection is mainly composed of digital scan documents (while most of our experts were not aware of this information). Note that the abbreviation in Table 7 corresponds to a concatenation of the two first columns of Appendix 2.

The other functionalities of text mining such as similarities search (search_sim: 0.90) and automatic classification of content (ana_content_class_text: 0.85) or search for NER (ana_content_ner: 0.80) were also of great interest. At the opposite, the two least popular functionalities, with a score of 0.30, were anonymization (ana_content_anonym) and readability-level evaluation (ana_content_readability). The complete list of the functionalities evaluation and their definition are given in Appendix 2 together with the metrics' evaluation.

### *Combination of data metrics and archival metrics*

*Appraisal tool mock-up.* An appraisal tool mock-up combining archival and data metrics was designed to illustrate the concept in a more concrete way. This model also allowed us to illustrate the functionalities during user evaluation. This result is presented in Figure 8 and is separated into seven areas.

**Table 5.**
Six main categories of functionalities

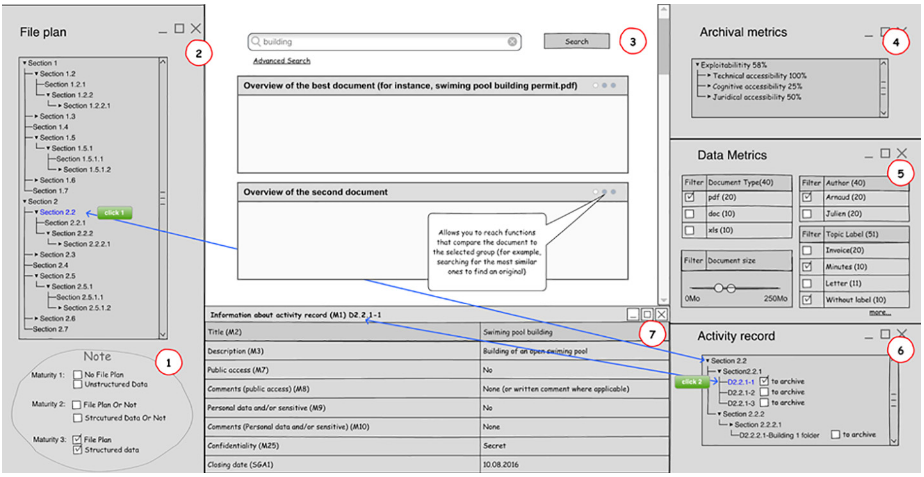
| Abbreviation | Description |
|---|---|
| preana_ | Document analysis occurring before the analysis |
| ana | Document analysis to extract new data from raw data |
| stat | Statistics on a group of documents |
| sear | Document search |
| learnauto | Machine learning approach from user interaction |
| admin | Administration of the appraisal tool |

**Table 6.**
Level of interest on the 33 functionalities for helping appraisal

| Level of interest | Average answer of interest (%) | Score interval |
|---|---|---|
| Particular interest | 70 | $0.8 \leq \text{score} \leq 1.0$ |
| Moderate interest | 24 | $0.5 \leq \text{score} < 0.8$ |
| Few or no interest | 6 | $0.0 \leq \text{score} < 0.5$ |

**Table 7.** Top ten functionalities selected by experts

| Abbreviation | Description | Score (range 0 to 1) |
|---|---|---|
| preana_index | Document extraction and indexing functionalities to retrieve metadata from file system raw data | 1.0 |
| ana_ocr | Optical character recognition functionality for text scanned document to treat them with all the text mining approaches | 1.0 |
| search_engine | Search functionalities to search in the documents any words in the text and the metadata | 1.0 |
| search_filter | Filter search functionalities on any metrics of the tool | 1.0 |
| preana_import_data | Import functionality for the raw data from a file system | 0.95 |
| preana_import_plan | Functionality to import a file plan in the tool | 0.95 |
| preana_import_records | Functionality to import the records (ISO 15 498) | 0.95 |
| ana_file_list_id_title | Functionality of file path analysis. Breakdown of the path into "title-identifier" when possible. This would make it possible to detect if a classification framework is in place, and if so, to link records (ISO 15 498) folders to section of the file plan | 0.95 |
| stat_time | Functionality to observe the number of documents created or modified during a period of time. Could be useful to create a story of document group | 0.95 |
| Admin_combo | Functionality of managing combinations of metrics | 0.95 |



Note

Maturity 1: □ No File Plan
□ Unstructured Data

Maturity 2: □ File Plan Or Not
□ Strcutured Data Or Not

Maturity 3: ☑ File Plan
☑ Structured data

Allows you to reach functions that compare the document to the selected group (for example, searching for the most similar ones to find an original)

Information about activity record (M1) D2.2.1-1

| | |
|---|---|
| Title (M2) | Swiming pool building |
| Description (M3) | Building of an open swiming pool |
| Public access (M7) | No |
| Comments (public access) (M8) | None (or written comment where applicable) |
| Personal data and/or sensitive (M9) | No |
| Comments (Personal data and/or sensitive) (M10) | None |
| Confidentiality (M25) | Secret |
| Closing date (SGA1) | 10.08.2016 |

| Filter | Document Type(40) |
|---|---|
| ☑ | pdf (20) |
| □ | doc (10) |
| □ | xls (10) |

| Filter | Document size |
|---|---|

| Filter | Author (40) |
|---|---|
| ☑ | Arnaud (20) |
| □ | Julien (20) |

| Filter | Topic Label (51) |
|---|---|
| □ | Invoice(20) |
| ☑ | Minutes (10) |
| □ | Letter (11) |
| ☑ | Without label (10) |

D2.2.1-1 ☑ to archive
D2.2.1-2 □ to archive
D2.2.1-3 □ to archive
D2.2.2.1-Building 1 folder □ to archive

**Figure 8.** Mock-up for illustrating concepts and functionalities of an appraisal tool interface

The first area is a simple indication to visualize the degree of maturity (here Maturity level 3, with structured information available and a file plan). The second area allows navigation through the file plan. The third area is more like a document search engine using the global context as a filter and allowing the viewing of relevant documents. The fourth area is represented as contextual facets of archival metrics (with dynamically calculated scores) that can also be used as a dynamic filter, if necessary. The fifth area works like the previous one (contextual and filter) but presents the low-level metrics from the data analysis. The sixth area allows display and navigation within the real repositories of the records being assessed. And finally, the seventh area allows you to obtain contextual information related to the filing plan and the files of the filing system.

*Archival and raw data mapping*. Figure 9 presents the conceptual model of the combination of the archival and data metrics. Data metrics can come from two types of source, either raw data (via content analysis (content), file system information or metadata included in certain data formats (meta)) or external information (via an archiving plan (file plan), records (records) and via a digital records system (DRS)). This raw data can be combined to create higher-level metrics (e.g. combination of the different dates found in a file, date in the file system, date in the metadata of the data format, date in the content).

Mapping between data and archiving metrics can be done (when possible) by mapping rules presented in the following Table 8.

The rules expressed in pseudocode (here simplified) generally apply to the characteristics of a particular document and have been created for 21 archiving variables. The final scores selected are arbitrarily either 0 for a variable that does not meet the criteria, 1 for a variable that meets the minimum criteria or 0.5 for situations where certain criteria are met. Then, three archival variables, which are V5, V23 and V32, were put into practice for the PoC, as illustrated in Table 9 for specific records in the dataOAEN collection. Thus, for the variable V5 from the root directory, there are 13,179 files, of which 66.1% get a score of 1 and 30.8% get a score of 0.5. In total, the score for this variable and directory is, therefore, 0.81 ((66.1 + 30.8/2)/100). In particular, Table 10 illustrates the granularity of the approach because the total scores for variable V5 can then be given for the root directory (0.81) but also for each sub-folder of the records folder.

## Limitations and discussions

When carrying out this project, we were aware there were some limitations. The main challenges were as follows:

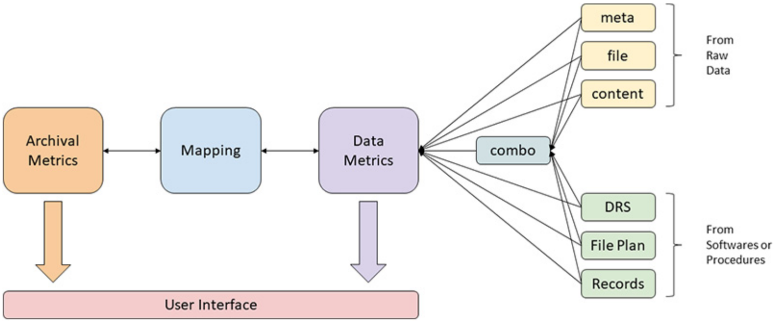


**Figure 9.**
Conceptual model for mapping archival and data metrics

**Table 8.**
Variable 5 (metadata completeness) mapping rules and corresponding score

| Mapping rule | Score<br>(range 0 to 1) |
|---|---|
| At least one date and one author in all metadata in the document | 1 |
| At least one date or author in all metadata in the document | 0.5 |
| No date or author metadata in the document | 0 |

- The interdisciplinary character and background of different members of research teams involved in the project. For that reason, we defined a glossary at the start of the project to clarify the terminology and to harmonize our terminological perception.
- The exploratory nature of the research itself was very challenging and the delimitation of the scope of our research was relatively difficult.
- As the scope was limited to the State of Neuchâtel only, there may be differences with other Swiss or other international contexts. However, the objective was to explore the feasibility of the automation of appraisal, even the effectiveness of the identified algorithmic approach still remains to be proven.
- The complexity of integrating structured data from different systems.
- Setting up an effective interface that can fit and provide answers to the needs of appraisal archivists.
- The limited sample size for the Likert test needs to be considered in the interpretation of results. Nevertheless, most of the identified 42 metrics have been subject to previous different validations. The latest were conducted among numerous tests in the State Archives of Wallis, under the Project QADEPs (2012-2013) (Makhlouf Shabou, 2014; 2015b).
- PoC is a good way to test ideas and operability. However, the relevancy of the approach and the computation of scores do still need to be evaluated seriously for most of the tested functionalities.

| | V5 | V23 | V32 |
|---|---|---|---|
| Archival variables | Metadata completeness | Nature of the file format | Presence of official languages |
| Directory path | Root/ | Root/1 SECTION | Root/2 SECTION |
| Number of files | 13,179 | 1,927 | 446 |
| Score of 1 | 66.1% | 64.9% | 53.1% |
| Score of 0.5 | 30.8% | 33.2% | – |
| Score of 0 | 3,1% | 1.9% | 46.9% |
| Total score | 0.81 | 0.81 | 0.53 |

**Table 9.** Synthetic scoring for three archival metrics based on data metrics on one top folder and sub-folders

| Directory path | Total score (range 0 to 1) | No. of files |
|---|---|---|
| D:\Data\OAEN-Lecteur_P | 0.81 | 13,179 |
| D:\Data\OAEN-Lecteur_P\0 SECTION | 0.76 | 2,076 |
| D:\Data\OAEN-Lecteur_P\0 SECTION\01 SUB_SECTION | 0.78 | 148 |
| D:\Data\OAEN-Lecteur_P\0 SECTION\01SUB_SECTION\011 SUB_SUB_SECTION | 0.78 | 148 |
| D:\Data\OAEN-Lecteur_P\0 SECTION\01 BASES LEGALES \011 LEGISLATION CANTONALE\011.1 LOIS | 0.77 | 107 |
| D:\Data\OAEN-Lecteur_P\0 DIRECTION ET COORDINATION\01 BASES LEGALES\011 LEGISLATION CANTONALE\011.1 LOIS\011.1-LArch2011 Loi sur l'archivage | 0.73 | 80 |

**Table 10.** Variable 5 score average for some directories of the file system of dataOAEN collection

## Conclusion

The objective of the study was to propose an appraisal tool to support decision-making. The study proposes an innovative and modular approach, because the measurement of ADs does not have to include all the variables, but can be based on a sample depending on the degree of maturity of the organization or its environment. It is adaptable because the folder creation context or the type of documents and metadata may require an adaptation of the measure and its weighting. Lastly, the possibility to reproduce, partially or exhaustively, the measurement of ADs systematically and accurately on a set of files remains a guarantee of efficiency and quality of our approach. In addition, this approach takes into account different documentary realities: situations where archives are structured and processed according to standards using automatic tools or even situations where archives, data and information are not processed. The potential facilities that artificial intelligence may bring to archival processing remain a real and effective contribution not only for day-to-day archivist's work but also for corporate information governance and decision makers.

## References

Belovari, S. (2018), "Expedited digital appraisal for regular archivists: an MPLP-type approach", *Journal of Archival Organization*, Vol. 14 Nos 1/2, pp. 55-77, available at: https://doi.org/10.1080/15332748.2018.1503014 (accessed 30 August 2019).

Belovari, S. (2019), "Simple and expedited digital appraisal/processing: two projects with the German state archives, Ludwigsburg (2016-2018)", paper presented at International Council on Archives Section on University and Research Institution Archives (ICA/SUV), Appraisal in University and Research Institution Archives, 1-3 July 2019, Dundee, Scotland, available at: www.ica.org/sites/default/files/icasuv2019_belovari.pdf (accessed 30 August 2019).

Harvey, R. and Thompson, D. (2010), "Automating the appraisal of digital materials", *Library Hi Tech*, Vol. 28 No. 2, pp. 313-322, available at: https://doi.org/10.1108/07378831011047703

International Organization for Standardization (2016), *ISO 15489-1: Information and Documentation – Records Management – Part 1: Concepts and Principles*, International Organization for Standardization, Geneva.

International Organization for Standardization (2018), *ISO/TR 21946: Information and Documentation – Appraisal for Managing Records*, International Organization for Standardization, Geneva.

Lee, C.A. (2018), "Computer-Assisted appraisal and selection of archival materials", in: *2018 IEEE International Conference on Big Data (Big Data)*, 10-13 December 2018, Seattle, WA, pp. 2721-2724, available at: https://doi.org/10.1109/BigData.2018.8622267

*Loi sur l'archivage* (2011), "(LArch; 442.20), 22 February 2011, state of Neuchâtel", available at: http://rsn.ne.ch/DATA/program/books/20132/pdf/44220.pdf.

Makhlouf Shabou, B. (2011a), *Étude Sur la Définition et la Mesure Des Qualités Des Archives Définitives Issues D'une Évaluation, École de Bibliothéconomie et Des Sciences de L'information*, Montréal, QC, available at: https://papyrus.bib.umontreal.ca/xmlui/handle/1866/4955 (accessed 29 August 2019).

Makhlouf Shabou, B. (2011b), "Measuring the quality of records to improve organizational documentary testimony", *2011 IEEE International Professional Communication Conference, 17-19 October 2011*, Cincinnati, OH, available at: https://doi.org/10.1109/IPCC.2011.6087223

Makhlouf Shabou, B. (2014), "Le projet QADEPs: un outil au service de la pérennisation des archives publiques", in: Hiraux, F., Mirguet, F., *De la Préservation à la Conservation: stratégies Pratiques D'archivage*, Academia l'Harmattan, Louvain-la-Neuve, pp. 87-98.

Makhlouf Shabou, B. (2015a), "Fonctions d'évaluation des archives: bilan sommaire des développements, des enjeux actuels et des défis futurs", in: Gagnon-Arguin, L., Marcel Lajeunesse, M. (Eds.), *Panorama de L'archivistique Contemporaine. Évolution de la Discipline et* *de la Profession. Mélanges Offerts à Carol Couture*, Presses Universitaires du Québec, QC, pp. 195-214.

Makhlouf Shabou, B. (2015b), "Digital diplomatics and measurement of electronic public data qualities: what lessons should be learned?", *Records Management Journal*, Vol. 25 No. 1, pp. 56-77, available at: www.emeraldinsight.com/doi/full/10.1108/RMJ-01-2015-0006 (accessed 29 August 2019).

Makhlouf Shabou, B. (2019), "On the path to an intelligent appraisal: innovative Swiss projects", paper presented at International Council on Archives Section on University and Research Institution Archives (ICA/SUV), Appraisal in University and Research Institution Archives, 1-3 July 2019, Dundee, Scotland, available at: ica.org/sites/default/files/icasuv2019_makhloufshabou.pdf (accessed 30 August 2019).

Makhlouf Shabou, B., Mellifluo, L. and Rey, R. (2013), *QADEPs: Définition et Mesure Des Qualités Des Archives et Documents Électroniques Publics*, Geneva School of Business Administration, Geneva.

Penn, E. (2019), "Appraising digital records, or swimming in treacherous shoals", paper presented at International Council on Archives Section on University and Research Institution Archives (ICA/SUV), Appraisal in University and Research Institution Archives, 1-3 July 2019, Dundee, Scotland, available at: www.ica.org/sites/default/files/icasuv2019_penn.pdf (accessed 30 August 2019).

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T. and Stuart, K. (2019), "More human than human? Artificial intelligence in the archive", *Archives and Manuscripts*, Vol. 47 No. 2, pp. 179-203, available at: https://doi.org/10.1080/01576895.2018.1502088 (accessed 30 August 2019).

Vellino, A. and Alberts, I. (2016), "Assisting the appraisal of e-mail records with automatic classification", *Records Management Journal*, Vol. 26 No. 3, pp. 293-313, available at: https://doi.org/10.1108/RMJ-02-2016-0006 (accessed 30 August 2019).

## Further reading

Archifiltre (2020), "Official website", available at: https://archifiltre.github.io/

International Organization for Standardization (1997), *ISO/IEC 15498: Information Technology — Data Interchange on 90 mm Optical Disk Cartridges — HS-1 Format — Capacity: 650 Mbytes per Cartridge*, International Organization for Standardization, Geneva.

Neuchâtel State Archives (2020), "Projet AENeas, state of neuchâtel's website", available at: www.ne.ch/autorites/DJSC/SCNE/archives-etat/numerique/Pages/73-ProjetAENeas.aspx

## Corresponding author

Basma Makhlouf Shabou can be contacted at: basma.makhlouf-shabou@hesge.ch

## Appendix 1

**Table A1.**
AD levels and related metrics

| AD level 1 | AD level 2 | AD level 3 | Variables | Related metrics |
|---|---|---|---|---|
| Trustworthiness | Reliability | Traceability of movements and operations | V1. Documentation of the transmission | Does information make it possible to track the transmission of the record to the archival entity? |
| | | | V2. Recording of events and actions related to document | In the case where interventions have taken place on the record during its life cycle, does information allow events to be traced? |
| | | | V3. Integration of the record in a document repository/system | Does information allow the record to be linked to a document repository integrating the classification and/or life cycle management aspect? |
| | | Completeness | V4. Completeness of the components of the record or file | Does information determine the presence of all the required elements in the document or file? |
| | | | V5. Metadata completeness | Does information determine the presence of all the required metadata? |
| | | Legal, regulatory and administrative compliance | V6. Authorization of the producer | Does information determine the legal capacity of the producer to produce the document? |
| | | | V7. Compliance of activity processing procedures | Does information make it possible to link the record or file to the administrative, regulatory or legal basis on which it was created? |
| | Authenticity | Identity | V8. Knowledge of the producer | Does information identify and authenticate the producer of the record or file? |
| | | | V9. Existence of an identifier | Does an identifier allow the record or file to be uniquely identified? |
| | | | V10. Existence of a title | Can a title be attached to the record or file? |
| | | | V11. File name format | Does the file name of the record or folder comply with the institution's internal naming rules? |
| | | | V12. Indication of creation dates | Is there any information to identify the date(s) of creation/validation of the record or file? |
| | | | V13. Existence of a signature | Does a signature authenticate the producer of the record or file? |
| | | Integrity | V14. Fixity of the bit string | Does a hash value ensure that the record or file has not been modified in an undocumented manner between its validation and consultation? |

(*continued*)

Table A1.

| AD level 1 | AD level 2 | AD level 3 | Variables | Related metrics |
|---|---|---|---|---|
| | Historical evidence | Traceability of activities | V15 Registration of the record in an official activity | Does information link the record or file to an official activity of the organization? |
| | | Rarity of testimony | V16. Exclusivity of information | Does information allow us to assess the scarcity of the information contained in the record or file compared to other sources of information? |
| | | Scope of the testimony | V17. Correspondence of the themes to a documentary repository | Does the record or file contain themes identified as particularly interesting? |
| | | | V18. Geographical relevance | Is there any information that links the record or file to a specific geographical area? |
| | | | V19. Temporal relevance | Is there information to link the record or file to a specific time area? |
| | | | V20. Frequency of activities | Does any information provide information on the frequency of the activity(ies) associated with the record or file? |
| Exploitability | Technical accessibility | Access functionality | V21. Validity of access paths | Is there any information to determine if the access paths are valid? |
| | | Functionality and rendering | V22. Significant characteristic | Does information provide information on the essential characteristics of the record that the institution must preserve? |
| | | Readability | V23. Nature of the file format | Is there any information that provides information about the file format of the record? |
| | | | V24. Creation application | Is there any information to inform on the application that created the document? |
| | | | V25. Restitution software environment | Does an information make it possible to inform about the existence of software for reconstituting the information in the record? |
| | | | V26. Storage medium | Is there any information about the storage medium used before deposit? |
| | | Long-term sustainability of access keys | V27. Availability of object access codes | Is there any information to indicate the existence and availability of access codes to access the record? |

(*continued*)

**Table A1.**

| AD level 1 | AD level 2 | AD level 3 | Variables | Related metrics |
|---|---|---|---|---|
| | Cognitive accessibility | Findability | V28. Multiplicity of entry points | Does information provide information about the existence of a cognitive identification system? |
| | | Comprehensibility | V29. Description of the producer | Is there any information that informs about the producer of the record? |
| | | | V30. Description of the creative context | Is there any information that informs on the reason for creating the record? |
| | | | V31. Description of the content | Does information make it possible to understand the content of the record? |
| | | | V32. Presence of official languages | Is the record provided in one or more of the organization's official languages? |
| | | | V5. Metadata completeness | Does information determine the presence of all the required metadata? |
| | Juridical accessibility | Availability of information | V33. Intellectual protection | Does intellectual property protection restrict the dissemination and exploitation of the record? |
| | | | V34. Data protection and privacy | Does protection exist that restricts consultation, dissemination and exploitation of the record? |
| Representativeness | Organizational context | Significance of the producer | V35. Producer's position | Is there any information to locate the producer of the record? |
| | | Significance of the document | V36. Type of document | Is there information to determine the type of document? |
| | | | V37. Relationship with the producer's functions | Is there information to make out the function that created the record? |
| | | | V38. Dissemination of the document | Is there any information to determine the circulation of the record? |
| | | | V39. Visibility of the producer network in the document | Does information make it possible to testify to the functioning of the organization that produces the record, particularly in its relational dimension? |
| | Socio-cultural context | Contextual rarity | V40. Originality of the context | Does information make it possible to distinguish the originality of the context in which the record was produced? |
| | | Aesthetic value | V41. Aesthetic quality | Does the record have any aesthetic value? |
| | | Heritage value | V42. Heritage quality | Does the record have heritage value? |

## Appendix 2

| Id | Type | Name | Description | Average score of interest (range 0 to 1) |
|---|---|---|---|---|
| 1 | preana | import_data | Functionality to import raw data from a file system in the tool | 0.95 |
| 2 | | index | Document extraction and indexing functionalities to retrieve metadata from file system raw data | 1.00 |
| 3 | | import_plan | Functionality to import a file plan in the tool | 0.95 |
| 4 | | import_records | Functionality to import the records (ISO 15 498) in the tool | 0.95 |
| 5 | ana | content_anonym | Functionality to anonymize the sensitive content (name, email) | 0.30 |
| 6 | | meta_sign | Functionality to recognize digital signature from metadata of a file | 0.75 |
| 7 | | file_list_id_title | Functionality of file path analysis. Breakdown of the path into "title-identifier" when possible. This would make it possible to detect if a classification framework is in place, and if so, to link records (ISO 15 498) folders to the file plan section | 0.95 |
| 8 | | content_ner | Functionality of NER to identify dates, names, locations, emails from text content | 0.80 |
| 9 | | content_class_image | Functionality of image classification to detect hand signature, official stamp, etc. | 0.75 |
| 10 | | content_class_text | Functionality of text classification to identify some type of content such as minutes, copyright, etc. | 0.85 |
| 11 | | content_detect_lang | Functionality of language detection | 0.65 |
| 12 | | content_summary | Functionality of automatic summarization | 0.60 |
| 13 | | content_readability | Functionality to attribute a score of readability (easy to hard to read, such as Flesch Kindcaid or Fog) | 0.30 |
| 14 | | content_link_record_plan | Functionality to create a link between file plan and records | 0.80 |
| 15 | | combo | Functionality to compose combination of data metrics | 0.95 |
| 16 | | metric_archiv | Functionality to compute archival metrics | 0.90 |

(*continued*)

**Table A2.** Data mining functionalities

| Id | Type | Name | Description | Average score of interest (range 0 to 1) |
|---|---|---|---|---|
| 17 | | ocr | Optical character recognition functionality for text scanned document to treat them with all the text mining approaches | 1.00 |
| 18 | | trans_image | Functionality to describe automatically an image | 0.65 |
| 19 | | trans_sound | Functionality of speech to text for audio and video content | 0.75 |
| 20 | | name_rules | Functionality to detect naming rules of file and directory | 0.80 |
| 21 | stat | count | Functionality to count any metric of a set of documents | 0.90 |
| 22 | | words | Functionality to extract frequent and relevant words | 0.80 |
| 23 | | time | Functionality to see the number of documents created and/or modified over time | 0.95 |
| 24 | | size | Functionality to count the total size of a document group | 0.85 |
| 25 | search | engine | Functionalities to search in the documents any words in the text and the metadata | 1.00 |
| 26 | | filter | Functionalities for filtering search with any metrics of the tool | 1.00 |
| 27 | | sim | Functionality to search for similar documents in a collection from one document or a group | 0.90 |
| 28 | | cluster | Functionality to create several clusters of documents from a query (unsupervised classification). Could be used to identify group of documents without any previous information | 0.80 |
| 29 | learnauto | text_gen | Text generation functionality: can generate titles, records file descriptions when they are missing | 0.75 |
| 30 | | class | Classification functionality for some metrics that may be missing, for example, proposal of a proposed final state, type of sampling, etc. when missing | 0.75 |
| 31 | admin | sample | Functionality to make/prepare a sample for archival purposes | 0.90 |
| 32 | | profile | Profile management functionality. This feature offers a more personalized tool linked to certain user preferences | 0.90 |
| 33 | | combo | Functionality of managing metric combinations and mapping with the archival model | 0.95 |

**Table A2.**