Collaboration and the Knowledge Economy: Issues, Applications, Case Studies
Paul Cunningham and Miriam Cunningham (Eds)
IOS Press, 2008 Amsterdam
ISBN 978–1–58603–924-0

Designing an Automated Prototype Tool for Preservation Quality Metadata Extraction for Ingest into Digital Repository

Milena DOBREVA[1,2], Yunhyong KIM[3], Seamus ROSS[3]
[1]*Center for Digital Library Research (CDLR), University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH, United Kingdom*
Tel: + 44 141 548 4753, Fax: +44 141 548 4523, Email: milena.dobreva@strath.ac.uk
[2]*Institute of Mathematics and informatics, 8 Acad. G. Bonchev St., Sofia, 1113, Bulgaria*
Tel: + 359 2 9792809, Fax: +359 2 9713649, Email: dobreva@math.bas.bg
[3]*Digital Curation Center (DCC) & Humanities Advanced Technology and Information, University of Glasgow, 11 University Gardens, Glasgow, G12 8QJ, UK,*
Tel: +44 141 330 5512, Fax: F: +44 141 330 3788,
Email: {y.kim|s.ross}@hatii.arts.gla.ac.uk

Abstract: We present a viable framework for the automated extraction of preservation quality metadata, which is adjusted to meet the needs of, ingest to digital repositories. It has three distinctive features: wide coverage, specialisation and emphasis on quality. Wide coverage is achieved through the use of a distributed system of tool repositories, which helps to implement it over a broad range of document object types. Specialisation is maintained through the selection of the most appropriate metadata extraction tool for each case based on the identification of the digital object genre. And quality is sustained by introducing control points at selected stages of the workflow of the system. The integration of these three features as components in the ingest of material into digital repositories is a defining step ahead in the current quest for improved management of digital resources.

1. Introduction

Institutions from different sectors (business, research, education and memory) seek to provide better management and organisation of objects in digital repositories. These repositories differ in aims, size, work procedures and approaches to their users. A study of the digital repositories for research publications carried out by the project DRIVER [1] and published in March 2007, shows that there are about 230 institutes with a digital repository for research output in the countries of the European Union. This report classified the countries in the EU into four groups with respect to the level of activity in establishing digital repositories for research publications: it found that in seven countries out of 27 no digital repositories exist or there is no evidence of existing repositories. Only seven out of 27 countries have well organised digital repositories with country-wide coverage.

As the repository development activities increase it seems paramount for the long term sustainability of these repositories to take a step back and re-examine the question of what methods are most efficient in populating repositories with digital objects that would have preservation quality metadata. It is essential to address the following observations [2]:

- Repositories which are at the beginning of their development need guidance;
- Even well-established repositories rely on manual collection of metadata;

- Manual collection of metadata results in a widely varying level of metadata quality across and within repositories because it is performed by actors with different background, capabilities, experience, expertise, and physical and emotional conditions;
- Manual collection of metadata is labour intensive and expensive.

Automated metadata generation can promote consistent quality across repositories and reduce the cost of collection. It could also lead to efficient means of building collection. The quality of metadata can further be enhanced by using a quality controlled modular approach to automation that employs a distributed management system. The methodological framework which we present here addresses all these issues. Our paper looks into these issues especially in the context of ingest of material into digital repositories, taking into account that sustainable management of digital repositories is dependent on identifying best practices in collecting high quality metadata to be integrated into its architecture from the very beginning, before or at the time of ingest.

2. Objectives

The paper aims to present ongoing research on ingest in repositories with an emphasis on pre-ingest activities featuring automated metadata extraction in a quality controlled environment. Then it suggests a preliminary framework for designing prototype tools for assisting preservation quality metadata extraction for ingest into a digital repository.

An additional objective of this paper is to highlight repository-related issues defined in the DigitalPreservationEurope Research Agenda[1]. It also seeks to raise awareness, amongst the specialists of the community, of future development needs, which would help to achieve excellence in digital repositories technology in the European landscape.

3. Methodology

We investigated previous approaches in extracting descriptive and semantic metadata automatically from digital material, and the tools and related research, which can be, incorporated into a general metadata extraction prototype tool.

The results of our investigations enabled us to define typical ingest workflows for digital repositories and what activities related to the evaluation of the quality of ingested digital objects and metadata should be inherent in the process. Based on our analysis of specific preservation quality metadata requirements and the ingest workflows, the task then examined what kinds of tools might increase both the capacity and quality of document ingest into digital repositories vis-à-vis automated metadata extraction.

We observed that current research and practical work in information extraction targeted at metadata are typically designed for very specific classes of documents (they extract metadata from documents of a specific genre and digital format) and lack generic applicability. The delivery of a generic tool, a ‘universal metadata extraction application’ capable of extracting metadata from any type of document without prior knowledge of the document type is not a realistic prospect—although it would be welcome. Even a brief consideration of the diversity of file formats, document types and structures, and domains in which information objects (e.g. documents) are created and used indicates the complexity of the problem. This observation has led us to adopt a different approach. We propose an approach that is based on the premise that, if we could determine the genre and technical format of a document, and if we knew what tools could be applied to metadata extraction from documents of the identified genre, we could ensure that the tool which would guarantee the highest recall and precision is selected and applied at the metadata extraction stage.

The genre of the document reflects essential properties relating to physical and conceptual structure of the document, i.e. genre classification clusters documents into

groups where each group consists of documents for which named metadata elements are likely to appear in relatively similar locations. We have already investigated the feasibility of automated genre classification including human labelling case studies and have found promising results [6, 7].

4. Technology Description

We propose a practical approach to genre classification which could be used in constructing automated metadata extraction workflow processes and would provide a foundation for the integration of supporting registries of classification and extraction tools with distributed workflow applications.

The model is founded on a consideration of a range of issues inherent in metadata extraction approaches: digital repositories, ingest workflows, automated metadata extraction, genre classification and quality issues. The resulting model proposes a novel approach: pre-ingest of automated metadata extraction based on genre classification of the digital object which allows choosing the most appropriate metadata extraction tool for selected genres. Among the innovative features of this model are that it foresees quality control, and it anticipates that metadata extraction tools will not necessarily be available for all classes of digital objects for which they will be needed and integrates this challenge into the workflow. This guarantees that only digital objects supplied with metadata of the desired quality would be ingested into the digital repository. The workflow provides a general framework for the automated extraction of preservation quality metadata for ingest into digital repositories. Since it follows a service-oriented distributed architecture, various tools within it could be implemented in different organisations and at different times (e.g. in response to specified requirements). To facilitate future implementations we also created a web registry of tools, which could contribute to different metadata extraction subtasks.

5. Developments

In the most popular existing workflows (see for example [3], [4], [5]), the digital objects presented for ingest arrive at the repository either accompanied by their metadata or have their metadata added after ingest. In both scenarios, mechanisms to support metadata quality control are lacking, and this poses risks to the long-term management of the digital objects themselves. Metadata quality has a lasting impact on discovery and retrieval, data and preservation management, and how future users can access the objects. The metadata extraction workflow described here is designed to be a pre-ingest process that includes quality control before the object is submitted to a repository. It is designed to ensure that digital objects ingested into a repository pass a metadata quality threshold; this threshold is defined at repository level.

To improve both quality and state of automation in digital repositories, we introduce automated metadata extraction into our workflow based on the assumption that an intelligent choice of an appropriate metadata extraction tool can be made according to the digital object format, genre and metadata quality requirements. The eventual deployment of a service based on this model depends upon the creation of a public repository of metadata extraction tools as well as the tools themselves.

The input into the workflow is generally a digital object of unidentified genre and format. This is received by the Digital Repository Content Manager, which is a process implemented by a software agent or a human user, or even by a combination of both, at various stages of the task. It initiates and guides the ingest process of digital objects into the repository and includes several transformations and decision-making points. The workflow implements the following core processes:

1. Digital object preparation, including digital object format detection and, if necessary, conversion to PDF.
2. Automated genre classification, involving analysis of the structure of the object and assignment of a genre.
3. Automated metadata extraction, featuring use of a distributed repository of metadata extraction tools for documents of various genres.
4. Quality control. A process where metadata are validated.

The input, output and repositories used in the four core activities are summarised below:

Table 1. Data flows in the automated metadata extraction workflow activities

Process	Data input	Data output	Repositories needed
1. *Digital object preparation*	Digital object	Digital object + Digital object in PDF	Repository of PDF converters
2. *Automated Genre Classification*	Digital object + Digital object in PDF	Digital object + Digital object in PDF + Genre	
3. *Automated Metadata Extraction*	Digital object + Digital object in PDF + Genre Format Quality Rights	Digital object + Digital object in PDF + Genre + Metadata **or** Request for a tool (when a metadata extraction tool does not exist)	Repository of automated metadata extraction tools Queue of digital objects of a genre for which a metadata extraction tool is not available
4. *Quality Control*	Quality requirements preset by	Ingest of digital object and metadata **or** repetition of the process.	Digital repository

The idea that we adopted in outlining the general architecture of the workflow was to encapsulate the separate processes described as independent *managers*. The architecture highlighting the managing component on the highest level is presented in Figure 1.

The output of the workflow depends on the outcome of the quality control with respect to the extracted metadata. In general, this outcome would be the document enriched with PDF representation, genre identification and metadata, ready for ingest into the repository. If metadata cannot be generated or do not meet the quality requirements, the process may be repeated. If the reason for the lack of metadata is the lack of availability of an appropriate metadata extraction tool, the digital object will be placed in a queue until the appropriate tool can be acquired (the workflow envisages communication with a public registry of metadata extraction tools).

6. Results

As mentioned above, Figure 1 presents the framework model on the highest level as a combination of processes and data flows. The 3D boxes present the five managers (PDF Conversion, Genre Classification, Metadata Extraction, Quality Control and Metadata Extraction Tools Manager). Decision points are represented by lozenge shapes. The central managing process is handled by the Digital Repository Content Manager (represented by the icon). We have used red dotted arrows to represent its intervention. The regular flow of activities is indicated by black arrows. The digital objects and other data generated by the

various processes are presented as blue and green rhombuses. The repositories used at various stages are also represented in Figure 1, as white document stacks with small icons.

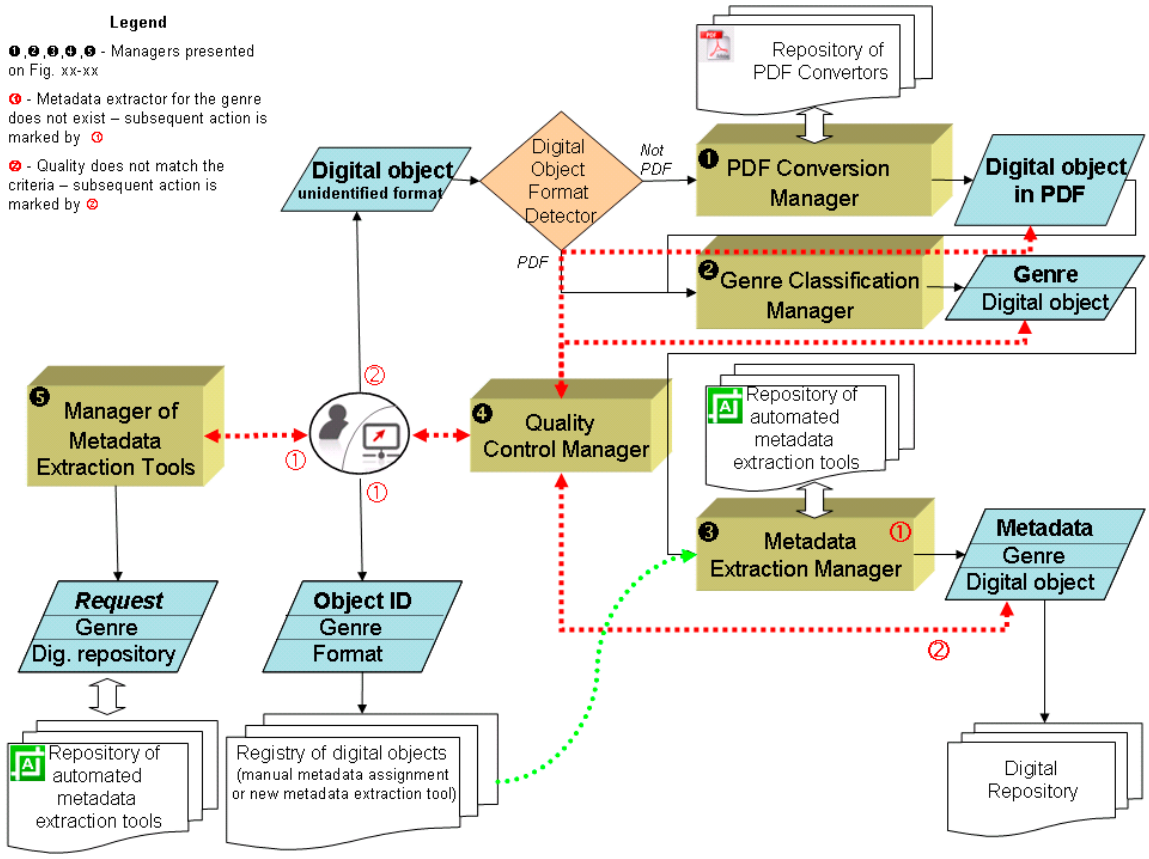

Fig. 1. Ingest framework

The workflow starts with submission of a digital object in an unidentified format for metadata extraction. Operation on the object is initiated by the Digital Repository Content Manager. For the sake of simplicity, we only consider here the situation where one object is processed at a time. Our assumption is that, in the case of multiple objects, the Digital Repository Content Manager will place them in a queue and that they will be processed consecutively.

When a digital object is presented for metadata extraction, the first step is to determine whether it is in PDF format. If it is PDF, then the object is submitted directly to the Genre Classification Manager. Otherwise, it will be submitted to the PDF Conversion Manager for analysis and representation of the object in PDF format. Conversion to PDF is intended to make all documents conform to one format for processing by the Genre Classification Manager, which we have optimised to work with PDF representations. The genre classification method, however, does not rely on PDF conversion, and can be modified to work with other formats in the future. The instance of the object in its original format is also retained and preserved and linked to the process. .

If the object format is not recognised as PDF, the first task of the PDF Conversion Manager is to identify the technical format (e.g. RTF, PS, JPEG) of the object. This is carried out within a Format Recognition Component. Once the format is identified, a tool is located to convert the object to PDF. The format influences the tools that will be needed to convert, render, and/or access the object. If the format is not known or a tool for the particular format is not available or does not exist, the Digital Repository Content Manager will decide how to proceed. A possible scenario might be to publish a public request for the necessary tool.

If a converter exists, it produces a PDF version of the digital object, which is checked for quality and sent to the Genre Classification Manager. When the Genre Classification Manager receives a PDF file, the process starts with an analysis performed by the Compound Object Handler to determine whether this is a simple or compound document. In the case of compound objects, it would create a queue consisting of the object followed by its sub-components. For example, in the case of books, journals or websites it is recommended to extract metadata not only on the higher-level genre but also on the constituent smaller identifiable pieces. For compound objects, the metadata extracted from the components will be integrated to form a composite metadata set at the end of the process. The Genre Classification Manager then proceeds to analyse the genre of the object and each of its components to label them with the genre to which they belong.

Each digital object is processed by a Submission Engine. Its role is to decide which classifiers to apply to a particular digital object. The model incorporates five classifiers: involving visual layout, language model (e.g. N-gram model of words), stylo-metrics (e.g. frequency of definite articles) and semantics of the text (e.g. number of subjective noun phrases), and domain knowledge (e.g. document source or format) [6]. In determining the genre of a digital object, these five classifiers are used discriminatingly, as not all features are necessarily expected to be present in the object, and the feature type most suitable for detecting documents of one genre is not necessarily the best for detecting documents of another genre [7]. Each of the classifiers applied will return a label value. If the classifier had not been used or could not extract any features from the object, it would return a null value, which is also informative in further analysis.

The acquired values are submitted to a Genre Labeller. Its decision-making tool uses an estimated probability distribution of features in relation to classes in a selected training data set to predict the genre class or classes of a document from a predefined schema. If agreement on the genre cannot be achieved, this tool communicates with the Digital Repository Content Manager, which would typically resubmit the object possibly with modifications for a new iteration of the genre-labelling exercise. The output of the tool is the digital object tagged with its genre label. The Quality Manager again takes the lead before the result (an agreed genre label) is submitted to the next component, the Metadata Extraction Manager.

The Metadata Extraction Manager deploys information gathered about the digital object and knowledge of its genre class to select the most appropriate metadata extraction tool from the Repository of Metadata Extraction Tools. Ross, Kim and Dobreva [2] have examined at least eleven research initiatives targeted at metadata extraction for documents belonging to specific genres (e.g. scientific articles or webpages). In selecting the metadata extractor, threshold settings for metadata depth and quality as defined by the Digital Repository Content Manager are taken into account.

A request for tools (from which to select the best available tool) consists of a set of values [g, f, r, q] constructed to represent Genre (g) and Format (f) described above, Quality (q) (described in the next paragraph), and Rights (r), where (r) is intended to convey the Digital Repository Content Manager's preference with respect to product license type (e.g. free or commercial) when selecting tools from the Repository of Metadata Extraction Tools. The Request Dispatcher then selects tools matching the values in the request. The most suitable metadata extractor is selected by submitting the retrieved tools to the Results Optimiser, which chooses the metadata extraction tool that has demonstrated greatest success on earlier occasions. If tools for a particular genre for either the PDF or the format in which the digital object was submitted are not available in the Repository of Metadata Extraction Tools, a check would be carried out to see what formats could be processed. The Metadata Extraction Manager could, as a result, initiate a process to generate a version of the digital object in a format that could be processed by an available metadata extraction

tool. After the digital object is submitted to the chosen metadata extractor quality control is applied to the extracted information before ingest. Should an appropriate tool not be available in the Repository of Metadata Extraction Tools the Manager of Metadata Extraction Tools handles the exception. The manager invokes the Request Initiator which starts an external search for an existing tool and if that fails to produce results announces to the community a request for open-source development of such a tool. Second, it adds information about the specific digital object and in which digital repository it is held to a registry of digital objects for which metadata extraction tools are unavailable.

The Quality Control Manager checks the results from the various managers against a predefined and repository-weighted set of quality parameters, including precision and recall, consistency, sufficiency, and trustworthiness: *Quality* threshold value (q). Quality parameters will be of indicative value only if metadata extraction tools have been tested against a transparent benchmark data set before their inclusion in the Repository of Automated Metadata Extraction Tools. This tool is fine-tuned by the Digital Repository Content Manager. It is the basic instrument for assuring the desired quality level. If the quality control leads to a positive result, the object is ingested into the repository or a set of repositories, some which may be distributed. Quality assurance processes would also be implemented for the results of PDF Conversion and Genre Classification Managers. If the quality control finds that the metadata do not pass the defined quality threshold, the digital object returns to the Genre Classification Manager with a request for its genre classification to be re-evaluated. As it would be pointless to return the digital object for genre re-assignment repeatedly after a certain number of failed attempts (although we have not yet identified the optimal number) the object will be sent to the Digital Repository Content Manager for further manual inspection.

There are two scenarios by which a digital object can be removed from the registry. The first depends upon manual extraction of metadata. The second is to return the digital object to the Manager of Metadata Extraction Tools when the necessary metadata extraction tool becomes available.

7. Business Benefits

We have proposed and described a metadata extraction workflow methodology that enables the community to pinpoint domains in which new research is needed and where new tools are required. We have suggested how services could be combined to incorporate automated metadata extraction into the ingest process for digital objects into a digital repository in a distributed environment. This will maximise the effectiveness of applying research results, and enable the exchange of tools developed in different institutions for different parts of the process.

The next step should include further refining the workflow, to elucidate better the processes, completing the survey of available tools and making it extensible by the community via an open-access tools repository, defining protocols and ‘application interfaces’ to support interoperability at the intersections within the framework, and deploying this conceptual model for practical testing by the community.

The continuation of this research and its implementation will also require additional studies of the user needs for the professionals in digital repository management. While the need of developing such tools is recognised, the diverse ways in which they can be integrated into real-life practices need further research and marketing.

8. Conclusions

The paper presents an approach to automating one particular process – the automated metadata extraction for and prior to the ingest of material into digital repositories. We have

suggested a model based on a distributed architecture supported by underlying quality control maximising the potential integration of specialised tools developed across different institutions. The benefits of the implementation of this approach will contribute to several critical components of digital repository management systems:

1. Higher amount of ingested digital objects supplied with preservation quality metadata contributes to sustainability of resources.
2. Improved quality of metadata provides a broader base for the retrieval components and should lead to higher user satisfaction.

Automation, especially in a quality-assured environment, is one of the areas of high demand for research and implementation work in the future years. This paper will help the specialists in digital libraries to understand better the current context of digital repositories and related research needs in realising automated processes. The general principle underlying the approach could be applied also for other repository-related activities.

Acknowledgements

This research has been supported under the DELOS: Network of Excellence on Digital Libraries (G038-507618) project funded under the European Union's Sixth Framework Programme. It also benefited from work being conducted as part of the Digital Curation Centre's (DCC) research programme.

References

[1] Van der Graaf, M.: Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU, White paper, version 0.9, March 2007, http://www.driver-support.eu/documents/DRIVER%20Inventory%20study%202007.pdf

[2] Ross, S., Kim, Y. and Dobreva, M.: Preliminary framework for designing prototype tools for assisting with preservation quality metadata extraction for ingest into digital repository, Pisa, DELOS NoE, December 2007, ISBN 2-912335-39-6.

[3] Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1 (2002), http://public.ccsds.org/publications/archive/650x0b1.pdf

[4] Bekaert, J. and Van de Sompel, H.: A Standards-based Solution for the Accurate Transfer of Digital Assets, D-Lib Magazine, Vol. 11(6) (2005), http://www.dlib.org/dlib/june05/bekaert/06bekaert.html

[5] Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G. and Smith, M.: The DSpace Institutional Digital Repository System: Current Functionality. Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, Houston, Texas, 27-31 May 2003. IEEE Computer Society, Washington, DC, pp. 87-97 (2003), http://ieeexplore.ieee.org/iel5/8569/27127/01204846.pdf

[6] Kim, Y. and Ross, S.: Genre classification in automated ingest and appraisal metadata. In: Gonzalo, J. (ed.): Proceedings of the European Conference on advanced technology and research in Digital Libraries, LNCS, Vol. 4172, pp. 63-74. Springer (2006), http://www.springerlink.com/content/2048x670g9863085/

[7] Kim, Y. and Ross, S.: Examining Variations of Prominent Features in Genre Classification. In Proceedings 41st Hawaiian International Conference on System Sciences, IEEE Computer Society Press, ISSN 1530-1605, (2008), http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=4438696&arnumber=4438835&count=502&index=138

[i] DigitalPreservationEurope: DPE Research Roadmap, DPE-D7.2, (2007), http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf