Automatic Preservation Watch using Information Extraction on the Web

A case study on semantic extraction of natural language for digital preservation

Luis Faria
KEEP SOLUTIONS
Rua Rosalvo de Almeida, 5
Braga, Portugal
lfaria@keep.pt

Alan Akbik
Technical University of Berlin
Einsteinufer, 17
Berlin, Germany
alan.akbik@tu-berlin.de

Barbara Sierman
National Library of the Netherlands
Prins Willem-Alexanderhof, 5
Den-Haag, Netherlands
barbara.sierman@kb.nl

Marcel Ras
National Library of the Netherlands
Prins Willem-Alexanderhof, 5
Den-Haag, Netherlands
marcel.ras@kb.nl

Miguel Ferreira
KEEP SOLUTIONS
Rua Rosalvo de Almeida, 5
Braga, Portugal
mferreira@keep.pt

José Carlos Ramalho
University of Minho
Braga, Portugal
jcr@di.uminho.pt

ABSTRACT

The ability to recognize when digital content is becoming endangered is essential for maintaining the long-term, continuous and authentic access to digital assets. To achieve this ability, knowledge about aspects of the world that might hinder the preservation of content is needed. However, the processes of gathering, managing and reasoning on knowledge can become manually infeasible when the volume and heterogeneity of content increases, multiplying the aspects to monitor. Automation of these processes is possible [11, 21], but its usefulness is limited by the data it is able to gather. Up to now, automatic digital preservation processes have been restricted to knowledge expressed in a machine understandable language, ignoring a plethora of data expressed in natural language, such as the DPC Technology Watch Reports, which could greatly contribute to the completeness and freshness of data about aspects of the world related to digital preservation.

This paper presents a real case scenario from the National Library of the Netherlands, where the monitoring of publishers and journals is needed. This knowledge is mostly represented in natural language on Web sites of the publishers and, therefore, is difficult to automatically monitor. In this paper, we demonstrate how we use information extraction technologies to find and extract machine readable information on publishers and journals for ingestion into automatic digital preservation watch tools. We show that the results of automatic semantic extraction are a good complement to existing knowledge bases on publishers [9, 20], finding newer and more complete data. We demonstrate the viability of the approach as an alternative or auxiliary method for automatically gathering information on preservation risks in digital content.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; H.3.7 [**Information Systems**]: Information Storage and Retrieval—*Digital Libraries*

Keywords

Digital preservation, monitoring, watch, natural language, information extraction

1. INTRODUCTION

Digital assets are continuously endangered by events that threaten user access or even cause irreparable loss of valuable content. These threats belong to many distinct domains, from technological to organizational, economical and political, and can relate to the holder of the content, the producers, the target communities to which the content is primarily destined for, or other internal or external influencers.

In digital preservation, watch (or monitoring) is a key capability that enables the early detection of these threats [5]. As the volume and heterogeneity of assets increases, it becomes infeasible to manually monitor all aspects of the world that may hinder their long-term access. Considering the scale of the problem, the automation of the watch process becomes a necessary step to ensure proper digital preservation.

But to automate the watch process, one needs data on aspects of the world that might afflict the preservation of digital content. This data relates to facts that directly or indirectly represent preservation risks or indicate the need for further analysis. Furthermore, to be able to automatically

reason on the data and infer preservation risks, this data needs to be expressed in machine readable language, i.e. in an explicit and formal specification [12] such as in a controlled vocabulary or an ontology.

Formally specified digital preservation related data can be found in existing repositories like file format registries and tool catalogues, but is commonly incomplete and outdated [24]. Worse, data about other domains that indirectly affect digital preservation, such as organizational or economical, or narrower domains that relate to a specific use case, is hard to find in a formally specified way.

Estimations on the amount of available machine readable data can be found in the Web of Data initiative [7]. Up until September 2011, the Web of Data contained 31 billion RDF triples and about 504 million links from a series of domains like Life sciences, Publications and Media [8]. In contrast, the amount of information contained in the Web is estimated to be orders of magnitude larger; In 2008 Google announced that they have parsed more than one trillion documents[1]. Furthermore, it is estimated that the Deep Web, i.e. the parts of the Web not indexed by search engines, is up to 550 times larger than the Surface Web [6].

However, the size of the Web of Data cannot be directly compared with the whole size of the Web because it is not measured in the same units. We may be able to extract hundreds of relevant RDF statements from the content of one document, but none from another. Nevertheless, for sake of comparison, if we hypothetically assume that, in average, from a document in the Web we could extract 100 RDF statements, than the Web of Data would be 100 thousand times smaller than the Surface Web and 55 million times smaller than the Deep Web.

Information Extraction technologies [23] present a way of making this wealth of unstructured information on the Web available to machine processing. They extract the meaning of information represented in natural language (and other formats) and express it in formally specified data. For preservation watch, the use of Information Extraction technologies has the potential of greatly improving the completeness and freshness of available data and thereby improving the accuracy and completeness of the watch process.

In this paper, we investigate the use of Information Extraction methods in the Web to assist preservation monitoring. We use a state-of-the-art Information Extraction system to gather information for a real case scenario from the National Library of The Netherlands, in which monitoring of scholarly journal publishers is needed. This large scale experiment is conducted as a proof-of-concept to assess the viability and the potential use of such an approach. We derive institutional policies and the requirements for preservation watch for the scenario and identify sources of information with formally specified data, such as manually created registries, for use in the automatic preservation watch processes.

We evaluate the results of the Information Extraction system against information contained in existing registries. The results strongly indicate that the proposed method can be used to automatically gather large amounts of high quality information from large collections of documents on the Web. Furthermore, the results indicate the viability of the proposed approach as an alternative or auxiliary method for automatically gathering information on preservation risks in digital content in order to keep automatic preservation watch more complete and up-to-date.

[1] `http://googleblog.blogspot.pt/2008/07/we-knew-web-was-big.html`

2. RELATED WORK

In this chapter, we give an overview on preservation policies, automatic preservation watch and the task of Information Extraction.

2.1 Preservation policies

Preservation policies are formulated by an organisation to guide the process of preserving digital information. The SCAPE project[2] is investigating ways of specifying preservation policies in a machine readable format. Based on several guidelines that support the creation of policies, like the OAIS model [14] and the Audit and Certification of Trustworthy Digital Repositories [13] we distinguish three levels of policies:

- the high level or guidance policies of the organization;
- the policies describing the approach the organization intends to take to achieve the goals that are phrased in the high level policies, called preservation procedure policies;
- the lowest level, on which the policies are described in more detail and focused on, for example, special collections and usage, called control policies.

The high level policies and the preservation procedure policies are "human readable". The control policies can be human readable, but need to have a machine readable version in order to be included in automatic preservation processes, such as automatic preservation watch and planning.

2.2 Automatic preservation watch

Automatic preservation watch is a systematic and automatic process of monitoring aspects of the world that might influence the preservation of digital content, revealing preservation risks or opportunities (such as hardware cost reductions). Preservation watch is usually initiated by policies that define constrains or goals that must be monitored from time to time. To be able to automatically watch these aspects of the world, sources of information for various domains must be identified and automatically harvested. This information is then merged into a central knowledge base that ensures information is well structured and normalized. Assessing this information allows one to find preservation risks and opportunities.

Scout: a preservation watch system

[2] `http://www.scape-project.eu`

Scout[3] is a semi-automatic preservation watch system that provides an ontological knowledge base to centralize all necessary information to detect preservation risks and opportunities [11]. It uses plug-ins to allow easy integration of new sources of information. The knowledge base can be easily browsed and triggers can be installed to automatically notify users of new risks and opportunities. Examples of such notification are: content fails to conform to defined policies, a format became obsolete or new tools able to render your content are available.

Different classes of information sources where identified to be integrated with Scout:

Format registries like PRONOM and UDFR, are online services with structured and formalized information on formats. They have high quality and relevant information for the digital preservation domain but commonly very incomplete. The PRONOM adaptor is already available on the Scout source code. Other generic-domain file format registries exist and are commonly more complete and up-to-date, but have less structured and formalized information.

Software catalogues are online services with information on tools that render, migrate, analyze and compare files of diverse file formats. This catalogues can be digital preservation specific, like the TOTEM[4] and the SCAPE Component Catalogue (under development), or generic-domain software catalogues in which information is not very structured and formalized.

Digital repositories have information on producer activity and the user community access preferences and problems, which certainly concern the repository owner. Also, when aggregated and viewed as a whole, this information can provide insight on the global tendencies and reveal *de facto* standards. A reference implementation adaptor is already available for the RODA repository[5].

Web Archives, like digital repositories, web archives can be of interest for the whole community. The content profile and the renderability problems [17] found with modern browsers can give important insights, and serve as a representative set, of the global internet content and trends. A reference implementation adaptor is already available for the renderability analysis of web archive systems.

Content profiles provide an aggregate view on the content characteristics and metadata, specially of the technical type. When applied to digital repositories and web archives, content profiling can provide precious information needed for preservation. If this information is shared with the community and viewed as a whole, it allows valuable insight on the wider state of curated content and can serve as an up-to-date indicator revealing technology and usage trends.

Experiments with tools, like migrators, assessing their behavior, reliability, completeness and quality are usually made as part of the preservation planning process. The outcome of these experiments can be of much interest to other users that are considering using that tools with similar objectives.

Policies. On-going work on formalizing preservation control policies and their relationship with organization strategies and goals will enable monitoring of some of the organizational objectives and check the repository compliance.

Simulation. Based on the data gathered by Scout, models of digital repositories can be created to predict the consequences of preservation actions [26], which allows the inclusion of forecasts into the knowledge base and be alerted of preservation risks before they actually happen.

Human knowledge. Some of the knowledge needed for digital preservation is still tacit or unstructured. Having humans as a source of information would allow the watch system to act as the central place for any kind of knowledge relevant for digital preservation to be gathered and formalized, even when there is no other specialized platform to support it.

Scout limitations

In Scout, the knowledge gathered from the information sources must be normalized and formally specified, i.e. machine readable. This requirement relates to the need for cross-relating different information and allow automatic reasoning on the knowledge to infer preservation risks and opportunities. Many of the identified information source classes have unstructured information. A possible solution is to require that humans, or crowd-sourcing, to be used to introduce the unstructured information available in the Web. Information extraction technologies could play a important part on automating, or reducing the human effort, of the introduction of unstructured information into the watch system.

2.3 Information extraction

Information Extraction (IE) is the task of extracting structured information from unstructured data such as natural language text. The goal of IE is to make information machine readable. Extracted information is often represented in the form of subject-predicate-object triples, where each triple is the *instance* of one semantic *relation*. The *predicate* indicates the relation to which the triple belongs, while the *subject* and *object* are the two entities between which the relation holds. IE methods use *extractors* that are either manually created or trained using machine learning methods to sieve through large volumes of text and distill for each relation a list of relation instances.

Pattern-based Information Extraction. IE methods often use a pattern matching approach where for each relation a set of *patterns* is defined that if encountered indicate a relation instance. The simplest example of such pattern-based information extraction might be a regular expression that finds e-mail addresses in Web pages. Patterns are often defined as *lexico-syntactic* patterns [1] that match natural language text; We illustrate this with the following example: Suppose we are interested in finding which companies have acquired other companies (such as Google buying Motorola for example). We refer to this common example relation as CompanyAcquisition. As lexico-syntactic pattern we may define "[X] acquired [Y]", where "[X]" and

[3] `http://openplanets.github.com/scout/`
[4] `http://keep-totem.co.uk`
[5] `http://www.roda-community.org`

"[Y]" are placeholders for the subject and the object entities respectively. If this pattern matches a statement in natural language text, a triple for the corresponsing relation is extracted. So, if the extractor encounters the sentence "*Google acquired Motorola in 2011.*", the relation instance CompanyAcquisition(Google, Motorola) [6] is extracted.

Challenges and limitations. As each relation is expressed in natural language text in a multitude of ways, one core challenge of Information Extraction methods is finding all patterns that belong to a specific relation. When aiming to extract relations from an open domain corpus such as the Web, this problem becomes more challenging as there may be a potentially unbounded number of relations, for each of which one extractor with a set of patterns must be defined. When interested only in information from a specific domain of the Web (such as digital content preservation), another challenge is to identify and gather relevant natural language text upon which Information Extraction is performed. Current IE methods are mostly limited to working on *explicit* statements in natural language text; reasoning or inferring knowledge from implicit statements is a topic of current research [16].

Level of supervision. A range of research investigates how to reduce the workload of manually defining patterns with machine learning mechanisms that require different amounts of supervision. Approaches range from supervised [19] or declarative [15] approaches, to weakly supervised [18] and unsupervised methods [4]. Supervised and declarative approaches generally produce high quality extractors, albeit at a cost in human effort, while unsupervised approaches are useful for information discovery.

Our approach. In this paper, we apply both a declarative [2] and an unsupervised [4, 3] approach to address a specific information need in the domain of digital content preservation. Our goal is to find relevant information on the Web. We discuss our system and how we address the above stated Information Extraction challenges in Section 3.

3. CASE STUDY

In this chapter, we describe a specific real world scenario in which large amounts of machine readable knowledge are needed. This scenario represents a use case that requires information to be as *complete* and as *up-to-date* as possible, and illustrates how such information can be found in natural language statements on the Web. We derive semantic relations from this use case and use a state-of-the-art Information Extraction pipeline to gather data from the Web and extract the required information from natural language text. We describe each step of the Information Extraction process and perform an analysis of the extraction results.

The purpose of this experiment is to execute a *proof-of-concept* on the idea of using Information Extraction methods to assist preservation monitoring. In particular, we wish to examine the following questions: What is the potential of using IE technologies to assist preservation monitoring? What are limitations and challenges? What are the prospects for future work in this field?

[6]Often, relation instances are denoted by first giving the predicate (i.e. relationship type) in camel case, and then the subject and object entities in brackets separated by a comma.

3.1 A real world scenario

As scholars have become increasingly reliant on electronic versions of scholarly journals, long-term preservation of these resources has become of major importance and a growing need for the library community. The shift to journal content that is digital, online and held remotely has challenged the responsibility that libraries have in ensuring the continuity of access to these materials. The National Library of The Netherlands (KB) was one of the very first cultural heritage institutions to become aware of the emerging importance of digital resources. As early as 1998 the KB concluded an agreement with the Dutch Publishers Association to extend the Dutch voluntary deposit scheme to off-line electronic publications, and in 1999 a tender was issued for the development of a long-term storage facility for electronic information resources. As no ready-made commercial products were available at the time, the KB embarked on a joint project with IBM to develop the Digital Information Archiving System (DIAS). With the establishment of the e-Depot the KB has created in 2002 the first solution to provide permanent access to scholarly information. This goes beyond the national depository role of the KB as it also preserves publications from international, academic publishers that do not have a clear country of origin. Originally, the e-Depot was designed to preserve the electronic publications of the Dutch publishers, in agreement with the Dutch voluntary deposit scheme. Some of the early archiving agreements were signed with major scientific publishers based in the Netherlands, such as Elsevier and Kluwer. As these are internationally operating publishers, the question soon arose how digital resources which are simultaneously published all over the world, fit into traditional national deposit schemes. The answer was simple: they do not. The KB decided that a new international framework would have to be developed to preserve digital publications for the long-term. As such a framework does not come to be overnight, the KB took a step by opening up its own e-Depot facilities to digital resources published by international publishers. Content for the e-Depot is delivered directly by scholarly publishers who have agreed to participate in the KB archiving service. As of June 2012, the e-Depot has preserved over 18 million journal articles.

The problem with e-journals

Today there are three leading archiving organizations agreed to act as last resort for e-journal content. Besides KB e-Depot, Portico[7] and CLOCKSS [8] are providing permanent access to this type of digital materials. All three are working very closely together and are involved in the Keepers registry which is a resource to address "who is looking after which e-journals, how, and what are the terms of access?"[9].

The next step for the KB is to position the international e-Depot as a European service, which guarantees permanent access to international, academic publications on a European level [22]. There is a danger that e-journals become

[7]`http://www.portico.org`
[8]`http://www.clockss.org`
[9]`http://www.thekeepers.org`

Table 1: Distribution of titles per publisher

% of titles	Publishers	Titles per publisher
40%	9	> 310
50%	21	> 132
60%	52	> 52
70%	143	> 16
75%	267	> 7
80%	569	> 3

Table 2: Publisher (P.) size distribution [10]

P. Size	# journals	% of P.	% of articles
very small	1-10	97%	30.9%
small	11-50	2%	14.6%
medium	51-250	0.32%	6.9%
large	250-1000	0.04%	6.2%
very large	> 1000	0.08%	41.4%
Total	**17.565**	**4.993**	**1.628.354**

"ephemeral" unless we take active steps to preserve the bits and bytes that increasingly represent our collective knowledge. Besides the threat of technical obsolescence there is the changing role of libraries. In the past, libraries have assumed preservation responsibility for materials they collect, while publishers have supplied the materials libraries need.

These well understood divisions of labour do not work in a digital environment and especially so when dealing with e-journals. So we need new models and organizations to ensure safe custody of these digital objects for future generations. The KB has invested in order to take its place within the research infrastructure at European level and the international e-Depot serves as a trustworthy digital archive for scholarly information for the European research community.

The scalability problem

To preserve scientific publications for future research we need to keep as much as possible. That means that the e-Depot needs to cover as much as e-journal titles as possible.

According to Ulrichsweb[10], there are over 35.000 peer reviewed journal titles within the academic realm. Over 65% of them, about 23.000, are online journals. According to EBSCO[11] there are over 5.000 publishers who are publishing 25.000 electronic journals. Yet another number comes from Web of Science[12]. This gives over 12.000 e-journals from 3.200 publishers. Looking more closely to these numbers we find out that the 100 largest publishing companies publish almost 70% of the available titles, as shown in the Table 1. So 80% of the available journal are provided by 569 publishers. Beyond that we find a huge long tail. According the the numbers of EBSCO again there are 466 publishers with two journals and 4.000 publishers with only one journal. A similar view is given by Scopus[13], the citation-index of Elsevier. In 2009 it counted almost 5.000 journal publishers in its database. 97% of them publishes 1-10 journals. This is, however, a significant part of the available journal articles, over 30%. In other words, the long tail is very large and in this we have to deal with a large amount of companies, as it is shown in Table 2.

The Tables 1 and 2 show that there is a great deal of concentration of journal titles with a small group of publishers. With 21 large publishers we cover 50% of the journal titles listed by EBSCO. But they also show that we have to face a huge long tail with 80% of the publishing companies publishing only one title. For the coverage of an e-journal archiving service like the KB e-Depot it is fairly doable to sign agreements with the largest publishing companies and ingest their content in the archive. But after that the real work begins, knowing also that each year over 1.5 million scientific articles are published.

Coverage

The international e-Depot was set up to be a service for the European research community to give access to scientific e-journals in case the university repositories or the publishers' platforms, which currently provide access, are no longer available or able to do this. The coverage of the journal titles to be archived is of most importance. Archival services have the aim to cover as many titles and articles as possible. Collections need to be complete. In practice, many situations can influence this completeness, like publishers getting out of business or journals changes between publishers. This happens very often and is a real problem, not only for archives, but certainly also for libraries, who are the subscription payers. The transfer of a title from one publisher to another itself is not the problem. The problem is in the administration of the transfer. Users, like libraries and archives, need to know when a title has been transferred and which publisher has taken over the title and under which conditions. The Transfer Code of Practice from the UK Serials Group gives a set of rules for transferring journal titles:

> The Transfer Code of Practice responds to the expressed needs of the scholarly journal community for consistent guidelines to help publishers ensure that journal content remains easily accessible by librarians and readers when there is a transfer between parties, and to ensure that the transfer process occurs with minimum disruption. The Code contains best practice guidelines for both the Transferring Publisher and the Receiving Publisher. Publishers are asked to endorse the Code, and to abide by its principles wherever it is commercially reasonable to do so. [25]

So the code exists to facilitate the users but, in the real world, this does not always work. Publishers do not follow these rules or do so very late. Administrative handling has no priority for a publisher and is only done months after the actual transfer. This is very problematic not only for the libraries using the subscription, but also for the archives who expect titles to be received from publishers. But after

[10] `http://www.ulrichsweb.com`
[11] `http://www.ebsco.com`
[12] `http://wokinfo.com`
[13] `http://www.scopus.com`

a transfer it suddenly ceases to receive the title any more. This hinders the coverage and completeness of the archive. It also brings along a great deal of work in finding out where the title has gone and who is the new publisher. So it takes time and work and it is a problem for coverage.

3.2 From the scenario to information sources

If we translate this description of the International e-Depot into polices, we can see that the high level aim is to create a complete collection of international scholarly e-journals for long-term preservation and access by acquiring these e-journals from the publishers in order to serve the European research community, in case the university repositories are no longer able to do this.

In order to achieve this, various preservation procedure policies need to be developed. The list of scholarly e-journals needs to be identified and the related publishers need to be contacted. Once the relationship is established via an agreement, regular monitoring needs to take place in order to be assured that changes can be dealt with and that the goal of "completeness of the journal collection" will be achieved. For this monitoring, detailed control policies will need to be established. The following list describes indications of the situations that can occur:

- Publisher A had journal J, is there a journal J provided by publisher A at time T_1?
- Publisher A had journal J_1 and the journal has been renamed to J_2 (i.e. has changed title or ISSN), is there a journal J_2 provided by publisher A at time T_1?
- Publisher A transferred journal J to publisher B, is there a journal J provided by publisher B at time T_1?
- Journal J has ceased to exist, what is its most recent issue?

In order to monitor these situations, the search results need to be filtered automatically, based on control policies. This experiment is limited to the investigation whether it is feasible to acquire relevant information from the Web and relevant registries. This relevant information relates to existing scientific journals, identified by title or ISSN, and journal-publisher relations that specify which publishers provide a certain journal. This information is manually maintained by registries within the e-Depot and also in other similar repositories and it is aggregated in the Keepers registry. But the information in the registries is only relative to the journals they collectively keep, being difficult to use to it to ascertain the completeness of the journal safeguarding. Furthermore, due to the manual processes involved and the lack of cooperation from the publishers, is incomplete and outdated. Nevertheless, publishers provide this information on their Web sites in natural language. So there is a possibility here for expanding and improving the available information by using information extraction technologies. In the following experiment we will focus on using information extraction technologies to gather information that would allows us to detect the first situation described above, whereas a journal J is provided by publisher A at time T_1.

3.3 Experiment

In our experiment, we aim to find a list of journal titles discussed in the Web, as well as a list of journal-publisher attributions in order to discover which publisher publishes which journal. The experiment consists of three steps; First we execute a *data acquisition and pre-processing step* that first gathers relevant natural language data from the Web. We then perform *relation discovery* on this data to mine frequent extraction patterns and gain an insight into the semantic content of the crawled corpus. We then assign patterns to the relations we wish to mine from the corpus and execute a *relation extraction* step to mine instances for each relation of interest.

Data Aquisition and Pre-Processing

The first phase of the experiment is a data acquisition and pre-processing pipeline. Its goal is to gather large amounts of natural language text from the Web that has a high probability of containing statements that relate to our use case.

We implemented a focused crawler to address this task. It uses a list of seed keywords (such as publisher names and journal titles) and formulates a search query using a Web search engine API[14] for each keyword on the list. Each query returns a list of Web pages that is automatically crawled and processed with Natural Language Processing (NLP) tools. Boilerplating is applied to extract blocks of natural language text from each Web page, removing other Web page elements such as layout information or advertisements. Sentence segmentation is applied to divide blocks of text up into sentences that can be analyzed individually. Finally, we filter out all sentences that do not contain at least one of the seed keywords.

The resulting dataset consists of approximately 18 million sentences gathered from 500.000 Web sites. The total text size is 8 GB. The seed keyword list consists of 12.000 entries. A sample of seed keywords and gathered sentences is illustrated in Table 3. These example sentences contain information relevant to our domain.

Relation Discovery

In the second phase of the experiment, we are interested to discover what kind of relations are expressed in the dataset. Because manually going through a dataset of this size is infeasible, we apply a relation discovery mechanism to identify prominent extraction *patterns* in the text. We apply a method explained in detail in [4] that counts and groups patterns according to distributional evidence in the corpus.

This yields a list of prominent patterns in the corpus, a sample of which is given in Table 4. These patterns indicate that the dataset gathered by the focused crawler is indeed relevant to our domain and suggests relationship types for which extractors can be created. Note that the patterns we use are actually more complicated lexico-syntactic patterns, but the syntactic elements (which denote grammatical properties of the patterns) are not indicated for the sake of readability.

Information Extraction

[14]In our pipeline, we use the Bing API, available at `http://www.bing.com/developers/` (last checked at 2013-04-20)

Table 3: Sample data from the data acquisition and pre-processing pipeline.

Seed keyword	Sample sentence retrieved from the Web
Elsevier	"*In 1991, two years before the merger with Reed, Elsevier acquired Pergamon Press in the UK.*"
The Asia-Europe Foundation	"*The Asia-Europe Foundation (ASEF) sold the Asia Europe Journal and transferred the copyright to its long-time partner Springer.*"
Acta Chirurgica Iugoslavica	"*Acta Chirurgica Iugoslavica is available free of charge as an Open Access journal on the Internet.*"
American Journal of Preventive Medicine	"*The American Journal of Preventive Medicine is the official journal of the American College of Preventive Medicine and the Association for Prevention Teaching and Research.*"
Journal of Business Ethics	"*In 2004 the Journal of Business Ethics merged with the International Journal of Value-Based Management and Teaching Business Ethics.*"

Table 4: Top pattern in the gathered corpus.

Pattern	Rank #
[X] journal of [Y]	1
[X] published by [Y]	2
[X] journal on [Y]	3
[X] journal published by [Y]	4
[X] available as [Y] journal	5
PubMed [X] [Y]	9
[X] science proceedings of [Y]	25
[X] subscription available to [Y]	30

In the third phase of the experiment, we wish to create extractors to find two relations in the crawled document collection: an extractor for journal titles (IsJournal) and an extractor for journal-publisher attributions (JournalPublisher). For each extractor, we manually go through the top patterns found in the relation discovery step and select patterns to use for relation extraction. For the JournalPublisher for example, we assign among others the patterns "[X] journal of [Y]" and "[X] journal published by [Y]".

We then execute both extractors on the document collection and store all found relation instances in two lists: One list of all journal titles found in the Web crawl, and one list of all identified journal-publisher attributions.

Information insertion into Scout

The resulting lists of journal titles and journal-publisher attributions conform to the formally specified and normalized information source restriction of Scout, the automatic preservation monitoring system explained in Section 2.2. This information can be inserted into Scout via a new plug-in, allowing this information to be included into the central knowledge base. Queries and notification triggers can then be created using the information on the knowledge base to alert when journals change publishers, or even to cross-relate an institution's list of subscribed publishers and journals of interest to alert when a journal of interest is no longer provided by any of the subscribed publishers.

The process of finding new journals and journal-publisher attributions used in this experience can be frequently repeated to allow automatic constant monitoring of these aspects of the world, automatically notifying interested users when the preservation risk of not acquiring a journal becomes relevant.

3.4 Results

In the experiment, we generate a list of 2,000 journal titles and a list of 500 journal-publisher attributions. We evaluated the results both automatically and manually against the e-Depot publishers. In the automatic evaluation, we matched the results against the e-Depot to find out how many of the extracted titles were already contained in the e-Depot internal registry[15]. Of the 2,000 journal titles, we found that only 200 were in the e-Depot, making the remaining 1,800 titles candidates for inclusion. We manually went through a sample of 200 of these titles and found that 191 are titles that should be added to the registry.

We manually repeated this experiment with the more complete Keepers Registry and found that more than 50% of all journal titles and 50% of all attributions were not in the registry and should be added. Again we found that the largest part of relation instances were viable candidates for entry into the registry. This indicates a strong potential of using Information Extraction technologies to help in keeping such registries complete and thus aiding the task of preservation monitoring.

In Table 5, we illustrate example instances of the JournalPublisher relation. The sample was chosen by sorting the list of all instances alphabetically by journal title and selecting the first 17 instances. The table illustrates which of these instances is already listed in Keepers Registry and which should be added to make it more complete. Some entries in the list have comments to illustrate error classes such as encoding errors or entity name boundary detection errors.

The information above can be directly used to answer the first situation described in section 3.2, whereas a journal J is provided by publisher A at time T_1, which is the time of data acquisition. The same IE pipeline can be frequently executed to get new snapshots in time, providing a continuous monitoring of this situation. Automatic monitoring of the continuity of e-journal availability can be done by cross-referencing this information about journal-publisher relations with the list of e-Depot paid publisher subscriptions (throughout time) and the list of e-journals available in the e-Depot repository. Nevertheless, for the other situations in section 3.2, more information about the journals needs

[15]the e-Depot archiving service contains an internal registry with the journal titles it archives and its related publishers, this internal registry is aggregated by the Keepers registry

Table 5: Sample of results and comparison to Keepers Registry.

Extracted relation instances		Evaluation	
Journal	**Publisher**	**In Keepers Registry?**	**Comment**
A Journal of Human Environment	Royal Swedish Academy of Sciences	**no**, should be added	
AAPS Journal	Springer Science + Business Media LLC	**yes**	
AAPS Journal	American Association of Pharmaceutical Scientists	**no**, should be added	
Academic Emergency Medicine	Society	**no**, error should be corrected and instance added	Error in entity detection of publisher name. It should be: "*Society of Academic Emergency Medicine*"
ACEEE International Journal of Network Security	ACEEE-Network Security Group	**no**, should be added	
Acta Applicandae Mathematicae	Springer	**yes**	
Acta Automatica Sinica	Chinese Association of Automation and Institute of Automation	**no**	"*Acta Automatica Sinica*" is listed only as published by "*Elsevier*".
ACTA AUTOMATICA SINICA	Chinese Association of Automation and Institute of Automation	**no**	All caps duplicate of previous relation
Acta Biomaterialia	Elsevier	**yes**	
Acta Geologica Slovaca	Comenius University in Bratislava	**yes**	
Acta Materialia	Elsevier	**yes**	
Acta Polytechnica Hungarica	âŤIJÃşâŢęÃëâŤijbuda University	**no**, error should be corrected and instance added	encoding error
Acta Radiologica	Scandinavian Society of Radiology	**no**, should be added	"*Acta Radiologica*" is listed with other publishers.
Aequationes Mathematicae	Birkh	**yes**	Encoding error in publisher name. Should be "*Birkhäuser Verlag*"
African Journal of Biomedical Research	Biomedical Communications Group	**no**, should be added	
Agricultural Economics	IAAE	**no**	"*Agricultural Economics*" is listed with other publishers. "*IAAE*" is an abbreviation for a missing publisher on the list.
Agronomy Journal	American Society of Agronomy	**yes**	

to be captured, like the journal renaming or ceasing. Also, machine readable information on the publisher subscription and e-journal issues available in the repository needs to be inserted into Scout in order to automatically cross-reference and discover other entailed preservation risks. These steps are some of the future work to be done in the Scape project to further study the use of information extraction technologies on digital preservation processes.

3.5 Lessons Learned

The experiment, intended as a proof-of-concept, strongly indicates the viability of using IE methods in preservation monitoring. We proceed to analyze the results more thoroughly with regards to sources of error (see Table 5) and potential improvements.

One major problem that affects extraction quality (i.e. the portion of results that are fully correct) is the lengthy nature of some journal titles or publisher names. An example of this is the "*European Journal of Nuclear Medicine and Molecular Imaging*". This causes our method to detect the wrong title boundaries in some cases; titles might be too short or too long, encompassing either only a portion of the words of the real title, or additional words that do not belong to it. We have adapted our method to cope with this, but more fine-tuning our extractors to this specific domain will arguably increase overall extraction quality.

We make another observation when we revisit the list of patterns in Table 4: We only used a small portion of the top patterns in our experiment. Incorporating additional patterns may lead to more complete extraction results. More importantly, we found that there were many types of information in the crawled corpora that were not extracted but may also be of interest to the community. For example, the pattern PUBMED [X] [Y] indicates that information on PubMed entries is contained in the corpus. Similarly, the pattern [X] JOURNAL ON [Y] indicates that it is possible to extract topics for journals. Accordingly, this indicates potential for expanding the range of information we extract in future experiments.

This experiment shows that the information extraction technologies has potential not only for detection of real-time threats for digital preservation domain, but also for parsing historical knowledge to capture descriptive information and becoming an important tool for librarians and archivists to cope with the increasing scale of digital content production.

4. CONCLUSIONS

Automatic preservation watch becomes a necessary capability of an institution when the factors that must be taken into consideration to do effective digital preservation become too complex or onerous for manual procedures. But automatic monitoring is highly dependent on the available machine readable information about the aspects of the world to monitor. Information extraction technologies can be used to surpass this limitation, allowing the use of information from the Web available in natural language.

The presented case study demonstrates how automatic monitoring can be done by using natural language statements from the Web. A real world scenario from the National Library of The Netherlands is presented where there is a need to monitor the scientific journal publishers, in order to ensure that there is an high coverage of all international scientific journals published throughout the world.

Sources for this kind of information are identified, like the Keepers registry and the e-Depot internal registry, but there are concerns that these registries may be incomplete and outdated. Information extraction technologies are then used to fetch natural language information dispersed throughout the Web and extract journal and journal-publisher attributions automatically. Comparing the information with the Keepers registry we find that more that 50% of the automatically fetched data is not on the registry and should be added, proving that this method is effective and can provide a much needed contribution for the automatic watch of the publisher community.

The technologies and methods used in the use case are not specific to publishing domain and can be applied to other monitoring needs, opening new possibilities for institutions to automate their watch processes. Using information extraction with automated preservation watch systems allows monitoring of non-technical domains, such as social, economical or organizational, where formally specified data is scarce. For example, monitoring economical or organizational changes in companies that support file formats or tools, like company bankruptcy or takeover, may allow the discovery of significant preservation risks. Also, this method allows monitoring of institutional specific domains, like the producer or target community, from which pre-existing formally specified data is rare and mostly manually created by institution itself. Further research on how to use these technologies and methods to monitor digital preservation related domains will be done in the next year of the SCAPE project.

5. ACKNOWLEDGEMENTS

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

6. REFERENCES

[1] A. Akbik and J. Bross. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In *Workshop on Semantic Search in Conjunction with the 18th Int. World Wide Web Conference*, 2009.

[2] A. Akbik, O. Konomi, and M. Melnikov. Propminer: A workflow for interactive information extraction and exploration using dependency trees. In *ACL System Demonstrations*. Association for Computational Linguistics, 2013.

[3] A. Akbik and A. Löser. Kraken: N-ary facts in open information extraction. In *AKBC-WEKEX*, pages 52–56. Association for Computational Linguistics, 2012.

[4] A. Akbik, L. Visengeriyeva, P. Herger, H. Hemsen, and A. Löser. Unsupervised discovery of relations and discriminative extraction patterns. In *Proceedings of the 24th International Conference on Computational Linguistics*, 2012.

[5] G. Antunes, J. Barateiro, C. Becker, J. Borbinha, D. Proença, and R. Vieira. Shaman reference architecture (version 3.0). Technical report, SHAMAN Project, 2011.

[6] M. Bergman. The deep web: Surfacing hidden value. *The journal of electronic publishing*, 7, 2001.

[7] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1265–1266, New York, NY, USA, 2008. ACM.

[8] C. Bizer, A. Jentzsch, and R. Cyganiak. State of the lod cloud. `http://lod-cloud.net/state/2011-09-19/`, 2011.

[9] P. Burnhill. Tales from the keepers registry: Serial issues about archiving & the web. *Serials Review*, 39(1):3–20, 2013.

[10] Elsevier. Scopus. `http://www.scopus.com`, 2009.

[11] L. Faria, P. Petrov, K. Duretec, C. Becker, M. Ferreira, and J. C. Ramalho. Design an architecture of a novel preservation watch system. In *International Conference on Asia-Pacific Digital Libraries (ICADL)*. Springer, 2012.

[12] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 2(5):199–220, 1993.

[13] ISO. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. ISO 16363:2012, International

Organization for Standardization, Geneva, Switzerland, 2012.

[14] ISO. Space data and information transfer systems - open archival information system (oais) - reference model. ISO 14721:2012, International Organization for Standardization, Geneva, Switzerland, 2012.

[15] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, and H. Zhu. Systemt: a system for declarative information extraction. *ACM SIGMOD Record*, 37(4):7–13, 2009.

[16] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[17] M. T. Law, N. Thome, S. Gançarski, and M. Cord. Structural and visual comparisons for web page archiving. pages 117–120. Proceedings of the 2012 ACM symposium on Document engineering, 2012.

[18] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[19] R. Mooney. Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 328–334, 1999.

[20] E. Oltmans and H. van Wijngaarden. Digital preservation in practice: the -depot at the koninklijke bibliotheek. *VINE*, 34:21–26, 2004.

[21] D. Pearson. AONS II: continuing the trend towards preservation software 'Nirvana'. In *Proc. of IPRES 2007*, 2007.

[22] M. Ras. The international e-depot to guarantee permanent access to scholarly publications. *Cultural Heritage On Line - Trusted Digital Repositories & Trusted Professionals*, 2012.

[23] S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272, 1999.

[24] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 1(6):165–182, June 2011.

[25] UKSG. Transfer. `http://www.uksg.org/transfer`.

[26] C. Weihs and A. Rauber. Simulating the effect of preservation actions on repository evolution. In *Proc. of iPRES 2011*, pages 62–69, Singapore, 2011.