Information Preservation in ARIADNE

6th DELOS Workshop – Preservation of Digital Information

Nuno Maria, Pedro Gaspar, António Ferreira, Mário J. Silva
{nmsm | pmag}@ui.icat.fc.ul.pt {asfe | mjs}@di.fc.ul.pt

ICAT/FCUL

Instituto de Ciência Aplicada e Tecnologia
Faculdade de Ciências da Universidade de Lisboa
Lisboa – Portugal

ABSTRACT

Preserving digital information is a necessary commitment to the future. In ARIADNE, a project of the Digital Publishing Group at ICAT, we are developing an integrated information system for news processing and publishing. In this paper, we present our perspectives on various information preservation issues addressed by this research. These include the semantic preservation of information classification schemes, preservation of the layout of dynamically generated documents, preservation of the linkage to external collections, and the economic sustainability of the news archive.

INTRODUCTION

Preserving or archiving digital information is, today, an important and necessary commitment to the future. How will we make available today's news to future readers? Giving the importance of this topic, the Digital Publishing Group of ICAT is studying new methods for preserving and processing heterogeneous information in organizations, combining multiple databases under a common framework. In project ARIADNE, jointly developed with *Público*, a national daily newspaper, we are building a new digital publishing structure, where all the information used and produced by journalists is organized in a common database (containing both data and metadata for collections maintained outside the organization). From the information in this digital library, we generate publications in digital format.

Público already maintains an archive of all the editions of its paper publications. This is a profitable unit within the company. The archive is used by the newspaper journalists and provides services to external entities. However, as we move into on-line publishing (we are about to release several new publications which will be available exclusively on-line) there is a need to define the processes for archiving and retrieving previously published on-line information. These new on-line publications differ significantly from the previous generation, where we were doing little more than creating on-line replicas of the paper editions. Our new publications are beginning to behave more and more as interactive user interfaces to databases of multimedia presentations.

In the next section we present the global architecture of our system, and then proceed with a more detailed discussion of the preservation issues in our digital library.

ARCHITECTURE

The architecture of ARIADNE is based on large multimedia data repository, which holds various collections of documents, newspaper articles, databases of readers and authors, places and events. For some collections, namely external publications, we only keep metadata and links for the articles.

The global architecture and the main information flow are shown in Figure 1. Several sources, including news agencies feeds, articles created for the paper edition of *Público* and external publications provide news items to the ARIADNE repository. Each article received is submitted to a preprocessing stage, where its metadata is extracted. The articles and news feeds are then converted into a common format based on XML[9], and archived in the collection repository with the software module *Loader*. Editions of electronic publications are built on a second stage, using another module, *Generator*, which selects a group of articles archived in the collections repository and packs them into presentations (or editions). This process finishes by converting the XML sources to HTML, making articles viewable from the current generation of web browsers. With this strategic approach we intend to overcome possible changes in data format standards and sustain our archival mission[3].

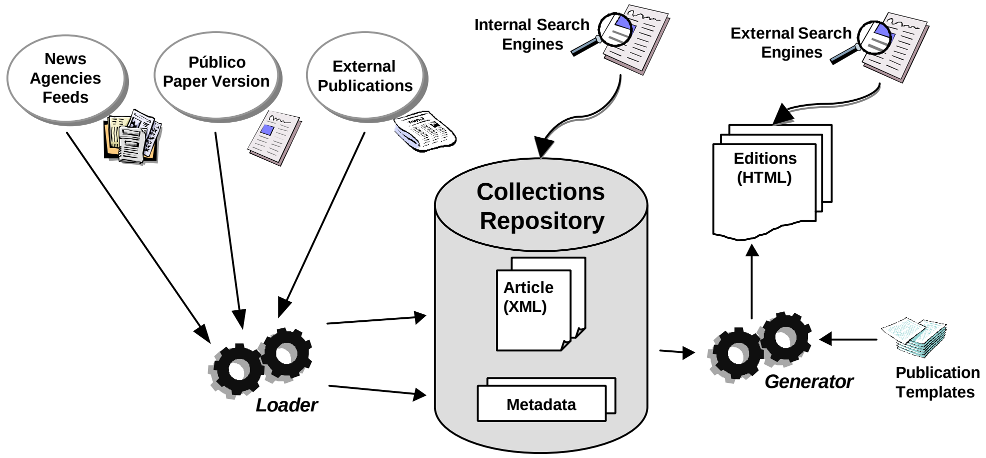

Figure 1. ARIADNE global architecture, and its main information flow. Each new article is classified, converted to XML and archived. Electronic publications are generated by picking articles from the various collections maintained in the repository.

Figure 2 shows the UML[8] class diagram of ARIADNE’s collections repository, with its main entities. The *article* represents the major information unit in this model. All other main classes are directly associated to it.

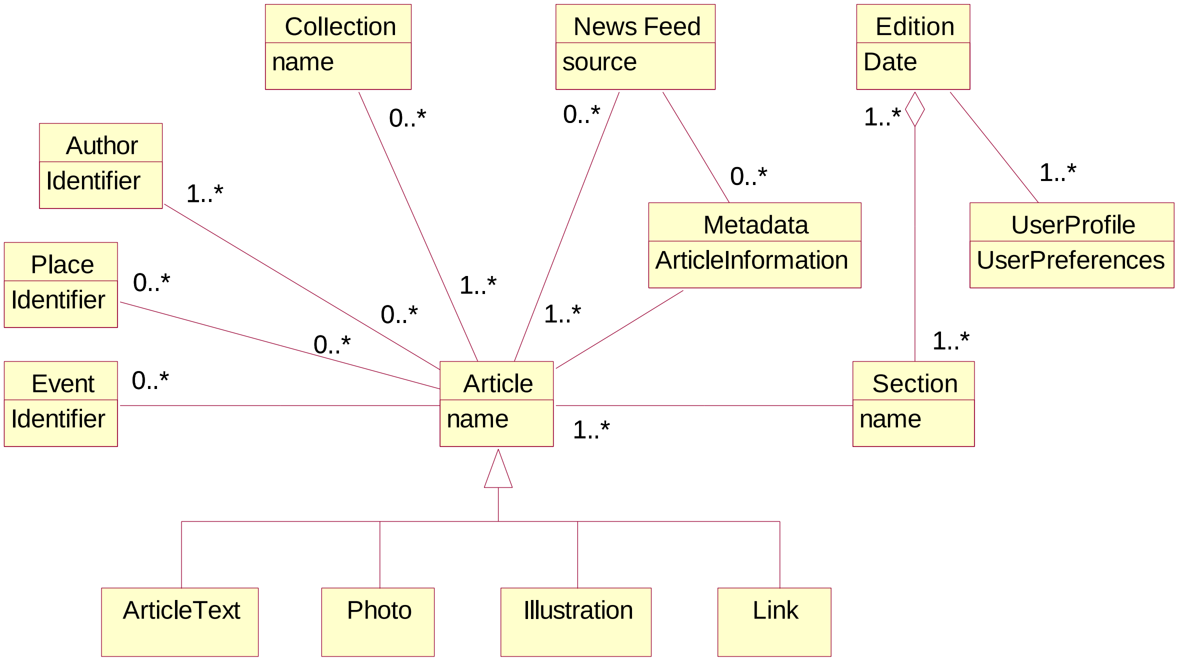

Figure 2. **UML model for the ARIADNE data repository architecture. The article is the major information object in this model. All other entities are directly related with this object.**

ARIADNE has also an internal search facility, combining information retrieval techniques with relational queries and data mining agents. Journalists and registered users will use this facility to retrieve information from the collections’ repository. As our publications are available on the World Wide Web, the main Internet search engines also index them.

Personalized publication is another major feature of ARIADNE. We use information-filtering techniques and maintain dynamic user profiles. These profiles are updated as a result of the data mining of access logs and user specified preferences. With this scheme, we intend to track users’ preferences as they change over time. However, this raises the problem for preserving personalized editions.

PRESERVATION ASPECTS

We face several problems associated with the archiving of the publications maintained in ARIADNE. First, we need to provide access to past editions of publications, with the additional complexity of preserving each of the personalized editions. As the personalized editions are dynamically generated from queries to the repository, we also need to preserve old information classification semantics when retrieving past editions.

Secondly, we need to preserve the layout of these personal editions. The layout format is one of the most characteristic aspects of a publication. The location and format of the articles within a page provide many visual clues, which are later used by readers to recall the same items.

Finally, as with all other digital libraries, we also face the problem of maintaining articles from external publications, handling their copyright restrictions, and recovering the cost of archiving the information.

The remaining of this section discusses our views on these preservation-related topics.

Preservation of Access Paths

Information classification is a complex, but necessary task for an easy discovery of resources. In ARIADNE, we use part of the work developed in the Dublin Core Workshop Series, and apply the 15 elements of the Dublin Core (DC) metadata element set[6] to each article in the data repository. In addition, we created complementary schemes for the specific domains of some of our collections. For instance, in our recipe collection, we added five additional descriptors: *recipe type, region of origin, preparation cost, difficulty and time*. This approach makes it easier and much more efficient to search and retrieve information in ARIADNE's data repository, but introduces the problem of preserving access paths to information items as the classification schemes and the retrieval interfaces change.

As concepts evolve over time, it is unrealistic to expect that the key characteristics of each information domain will remain static. In the previous example of our recipe collection, we started with a classification containing just four descriptors. Later on, *Público* decided to print several regional culinary books and we added the new *region of origin* attribute to the recipes loaded from these books. As the recipes already in the collection, were not classified with this descriptor, when a reader uses the new interface, with the new classification topic to select some regional recipes, these unclassified recipes will not be selected.

Another problem with this classification scheme raised when it was necessary to separate soup recipes from "*Açorda*" recipes (*Açorda* is a traditional Portuguese recipe made of soup and bread). Initially, both soup and *Açorda* recipes were classified only with the "soup" attribute, but the new retrieval interface distinguishes between these two kinds of recipes. This re-classification introduces a new problem: we now need to preserve the semantics of search results.

Information semantics based on information classification is a very important aspect in our architecture. We are studying in ARIADNE two different approaches to this problem. First option is to complement the retrieve mechanism, plugging two or more different components in the engine, which then will be responsible for the information retrieval in each single semantic classification domain. In this approach, the archive remains unchanged and there is no need for digital records reclassification. Another option would maintain a single and simple retrieve mechanism, but reclassify each digital record in the main repository according to the new semantic classification scheme.

Each solution has advantages and drawbacks. With the first, we have a scalability problem, as the introduction of new semantics implies adding new retrieve mechanism plugins, which delay retrieval operations. However, this solution has high maintainability because its centralized retrieval mechanism can be easily upgraded. With the second approach, we have a fast retrieval mechanism, but maintaining the collections will be hard, as we predict a huge, distributed data repository and the application of a conversion function for reclassification of a large number of digital records may be impracticable.

In the recipe collection, we fixed the "*Açorda*-soup" problem using the first approach, by inserting a plugin in the search engine that controls queries in this collection. The search engine understands each query and collects all related recipes, presenting them in the order implied by users' queries. Nevertheless a similar approach will not be efficient with the regional recipes' problem, because the criteria for identifying its region are not clear. In this case, a reclassification of the recipe collection would be a better choice.

Layout Preservation

How many times did we search for that particular second item on a report list, generated by a web search engine, even though we did not remember what was really written there? This shows the importance of preserving the order of presentation of information items within a digital document.

In a traditional repository of digital publications we might archive each edition with its user interface and it would remain forever that way. However, in ARIADNE, the existence of personalized editions raises the problem of also preserving the profiles that generate these editions. We use a method to archive each edition in a way that allows storage space consumption to increase only incrementally as we add new readers. In our architecture, a user profile maps a reader into one of a set of presentation styles (or templates). Publication editors define these styles to match typical profiles that they have identified among their readership. This approach still requires the preservation of the presentation styles used by each reader over time, so we can re-create the user interface and the layout of pages dynamically generated when a reader visits the same edition of a publication.

Technology Preservation

Another key topic in information preservation is the technology. We strive to make ARIADNE resist to continuous changes in representation formats, by adopting the latest standard and converting information into this format. When necessary, because some browsers may not yet support the new standard, we generate the information in the older formats. It may sound incoherent but, with this design, we can let readers access past editions, even if meanwhile they have migrated into new browsers, and we find it easier to convert data back into the original format when required. In ARIADNE we are already storing articles in XML, which is richer then HTML and will be directly viewable with the next generation of browsers. However, as almost our readers cannot read this format, we are currently generating editions in HTML.

We will keep on generating HTML editions until only a small fraction of our readers remains without the capability to process XML documents. It is a growing common ground the increasing conservatism of web users. Jakob Nielsen showed recently the increasing conservatism of web users, which frequently don’t have the latest client software available [5]. So it is important to keep ARIADNE’s publications easily accessible not only by the fifth generation Internet browsers but also by the first generation browsers. There is no point in converting information to the latest format if most of the readers cannot access it.

Linkage Preservation

It is impossible to us to track all of the journalistic information available on the web. We can not rely on available Internet search engines, as they do not index other newspapers of interest to us with the required frequency. As a result, we need to create our own index, and establish our own approach for preserving this linkage. For several of the external sites, we keep linkage and metadata information. This strategy overcomes several legal and scale problems, but it also raises other problems related to the preservation of the referenced contents. What will happen if the publisher changes information location or even ceases business [3]? How about the intellectual preservation of remote resources, there is no way to make sure the article maintains the integrity and authenticity of the information as originally recorded [1].

Facing these disadvantages we decided in some cases to collect entire external publications. As there is widespread uncertainty about legal requirements for managing intellectual property in digital environment[3], we restrict availability of this information to the journalists of Público. With this approach, we intend to increase our digital library functionality, avoiding corrupted information contents and outdated linkage.

Preservation Costs

The electronic archive being developed brings many costs and must sustain itself economically. Gillian Laughton presents an interesting comparative study of the archiving cost of an ASCII text based electronic journal with that of a new generation journal in HTML [4]. In an example of his study, once indexing and the increased staffing required to maintain the more complex formats are included, the overall costs jump from $435 to $1,000 per title per year. In our publishing system, access to recent news is free, and advertising is the only source of income. However, the use of all other services will be charged. These services include searches on our indexes and collections, and notifications sent by user created information agents. The services are about to be launched to the public and will be available on a subscription basis. We are starting the development of a micro-payments system, so that we can charge individual news articles. However, we do not have an idea of how much revenue these services will produce. This may be only be obtained through observation of the readers reaction to our pricing policies.

CONCLUSIONS

In ARIADNE we are concerned with some specific aspects of information preservation.

We attempted to describe in this paper our current concerns and our approach to building an archive of digital publications. This is becoming an increasingly harder problem to solve, as we move into personalized editions, which invoke dynamic queries and incorporate richer information types. However, in our view, the exact information available and its presentation organization at the time of publishing are essential aspects to preserve in the archive of a reference daily newspaper, such as *Público*.

A common concern of librarians is that computer science success is "in part because of its luxurious ignorance of the past", and this generation choose to delay the problem of archiving, " mainly because it is simply not as interesting as developing the wonders of future" [3]. In ARIADNE we are feeling the same pressure from our partner journalists.

REFERENCES

[1] Graham, Peter S., Preserving the Digital Library, Long Term Preservation of Electronic Materials Workshop, November 1995, <http://www.ukoln.ac.uk/services/elib/papers/other/preservation>

[2] Hall, Barbara, Archiving Electronic Journals, presented in the American Library Association Annual Conference, June 1997, <http://www-lib.usc.edu/Info/Acqui/research.html>;

[3] Issues and Innovations in Preserving Digital Information, in Transforming Libraries, Issue 5, ARL March 1998, <http://www.arl.org/transform/pdi>;

[4] Laughton, Gillian, Archiving of Electronic Journals, <http://solaris.cis.csiro.au/im/ejournal/archive.htm>;

[5] Nielsen, Jakob, The Increasing Conservatism of Web Users, March 1998, <http://www.useit.com/alertbox/980322.html>;

[6] Dublin Core Metadata, <http://purl.oclc.org/metadata/dublin_core>;

[7] HTML - HyperText Markup Language, <http://www.w3.org/MarkUp>;

[8] UML – Unified Modeling Language, <http://www.omg.org/library/schedule/Technology_Adoptions.htm#tbl_UML_Specification>

[9] XML Specification Version 1.0, December 1997, <http://www.w3.org/TR/PR-xml-971208>;